

The Musical Work as a Mutable Interface: Towards a Co-Creative Electroacoustic Practice

Nádia Carvalho



Supervisor

Gilberto Bernardes de Almeida

Prof. Dr.

Faculty of Engineering of Porto-INESC-TEC



The Musical Work as a Mutable Interface: Towards a Co-Creative Electroacoustic Practice

Nádia Carvalho

Doctoral Program in Digital Media

Approved by:

President: Dr. António Coelho

Referee: Dr. Diemo Schwarz

Referee: Dr. Rui Penha

Referee: Dr. Sofia Cavaco

Referee: Dr. António Sá Pinto

July 6, 2026

Resumo

Esta dissertação investiga a transição ontológica e performativa da obra musical, propondo a sua reconfiguração de um artefacto estático numa interface mutável através da mediação computacional inteligente. Historicamente, a tradição musical erudita ocidental privilegia o conceito de obra (*werktreue*). Este conceito cristaliza a composição como um objeto imutável, preservado pela rigidez da partitura ou do suporte fonográfico. Contudo, as práticas digitais e eletroacústicas contemporâneas reclamam novos paradigmas que respondam à porosidade das fronteiras entre composição, performance e agência tecnológica. Este trabalho propõe, assim, a substituição da partitura fixa pela variedade topológica navegável. Neste sentido, a obra redefine-se como um território dinâmico cuja identidade assenta na fluidez estrutural e não na prescrição notacional. Esta mudança de paradigma fundamenta-se num modelo de navegação contextual. Neste modelo, os limites topológicos do sistema e a intuição gestual do intérprete convergem num diálogo exploratório de descoberta estrutural em tempo real.

A metodologia adotada cruza a investigação experimental com a investigação baseada na prática. Nesse sentido, propomos uma estrutura em um ciclo reflexivo-iterativo que articula o desenvolvimento de software ao ato performativo. A investigação desenvolve-se em dois eixos fundamentais: a representação tonal simbólica e a navegação tímbrica. No primeiro, recorre-se a variational autoencoders (VAEs) para a compressão de polifonia simbólica, tendo-se observado um alinhamento empírico significativo entre os espaços latentes desses modelos e os espaços de fase da Transformada Discreta de Fourier (DFT). Este resultado demonstra que modelos considerados de caixa negra podem revelar, de forma autónoma, princípios estruturais da teoria musical, como o ciclo de quintas. No segundo eixo, utiliza-se a Síntese de Áudio Neural (RAVE) para mapear o espectro tímbrico do saxofone tenor em espaços latentes de alta dimensão. De forma a mediar a relação entre dados abstratos e o gesto musical. Para tal, introduz-se uma taxonomia de movimentos (paralelos, oblíquos e contrários) que preserva a agência do intérprete na navegação destes espaços.

A aplicabilidade prática desta investigação materializa-se nos sistemas BroadcastJSB e Aethra. O primeiro constitui um instrumento topológico que explora a metáfora do rádio para a navegação em tempo real no espaço latente dos corais de J.S. Bach. O segundo consiste numa interface co-criativa para música mista que sintetiza a complexidade multidimensional num único eixo de controlo performativo. As principais contribuições deste trabalho residem na formalização da obra mutável como espaço de possibilidades, na demonstração da inteligibilidade musical em modelos de aprendizagem profunda e na criação de ferramentas e repositórios de dados de acesso aberto às comunidades que investigam performance contemporânea e computação sonora e musical. Em suma, a dissertação sustenta que o futuro da prática musical reside na potencialização da intuição humana através da mediação inteligente, posicionando a interface mutável como o lugar primordial da exploração colaborativa na era digital.

Abstract

This thesis investigates the ontological and performative shift of the musical work. In this sense, we propose using intelligent computational mediation to reconfigure it from a static artefact into a mutable interface. Historically, the Western musical tradition has been framed by the work-concept (*werktreue*), which crystallizes the composition as an immutable object preserved by the rigidity of the score or the phonographic medium. However, contemporary digital and electroacoustic practices demand new paradigms that account for the fluid boundaries between composition, performance, and technological agency. Acknowledging that, this research proposes a transition from the static score to the navigable topological manifold. Rather than utilizing fixed musical notation, these spaces treat the musical work as a dynamic environment. In this environment, identity is preserved through structural fluidity, leading to a paradigm change to a contextual navigation model. In a symbiotic dialogue of real-time structural discovery, the machine provides a restricted, high-dimensional map of the work’s universe, which the performer navigates using their gestural intuition.

The methodology follows a mixed approach between experimental and practice-based research. This methodology operates over a reflective-iterative cycle, integrating software development with musical performance. The study is divided into two main phases: symbolic (tonal) and sub-symbolic (timbre) representation and navigation. In the first phase, variational autoencoders (VAEs) are used to compress symbolic polyphony. A critical milestone was the empirical alignment found between unsupervised VAE latent spaces and theoretical Discrete Fourier Transform (DFT) phase spaces. This alignment confirms that black box latent space models naturally discover structural principles consistent with established music theory, such as the circle of fifths. The second phase employs Neural Audio Synthesis (RAVE) to map the timbre landscape of the tenor saxophone into high-dimensional latent spaces. To navigate these spaces, the research introduces a taxonomy of musical motions (parallel, oblique, and contrary). These motions propose to connect abstract data and musical intent, enabling gestural navigation that honours performer agency.

The practical applicability of this research is materialized in BroadcastJSB and Aethra. The former, a web-based topological instrument, uses a radio metaphor for real-time navigation within J.S. Bach’s chorales’ latent space. The latter consists of a co-creative interface for mixed music that synthesizes multidimensional complexity into a single axis of performative control. The primary contribution of this work lies in the formalization of the mutable work as a space of possibilities. We also demonstrate musical intelligibility within deep learning models and create open-access tools and data repositories for the computer music research and contemporary performance communities. In summary, the thesis argues that the future of musical practice lies in improving human intuition through intelligent mediation. By doing this, we position the mutable interface as a key place for collaboration and creative exploration in today’s digital world.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Gilberto Bernardes, for his invaluable guidance, support, and the trust he placed in this work. His critical vision and constant encouragement were fundamental to the realization of this research, consistently challenging me to elevate both the scientific rigor and the artistic depth of this project.

I am especially grateful to Jorge Sousa, whose exceptional musicianship and technical expertise as a saxophonist were vital to the second phase of this research. His collaboration in the experiments and his willingness to explore new creative territories with the tenor saxophone provided the essential insights that allowed the theoretical models to come to life through performance.

I am also grateful to my host institution, INESC-TEC, for providing an environment of excellence and the technical and human resources necessary for the development of this work, as well as to all my colleagues who contributed to such a rich space for intellectual exchange.

This research was made possible by the financial support of the Foundation for Science and Technology (FCT), through the doctoral grant with reference [2021.05132.BD], to which I extend my sincere recognition.

Finally, a very special thank you to my parents for being my safe harbor throughout every stage of this journey. To my mother, in particular, whose unconditional love, patience, and constant support provided the driving force that allowed me to overcome the most uncertain moments. Without your example and dedication, this achievement would not have been possible.

Nádia Carvalho

“The poetics of the ‘open’ work tends to encourage ‘acts of conscious freedom’ on the part of the performer and place him at the focal point of a network of limitless interrelations, among which he chooses to set up his own form without being influenced by an external necessity which definitively prescribes the organization of the work in hand.”

Umberto Eco

Contents

I	Framing the Mutable Musical Work	1
1	Introduction	2
1.1	The Problem of the Fixed Work in Digital Practice	3
1.2	The Mutable Interface as a Research Framework	4
1.3	Contextual Navigation in Machine-Mediated Performance	6
1.4	Research Objectives, Questions, and Contributions	7
1.4.1	RQ1: On the Ontology of the Mutable Work	7
1.4.2	RQ2: On Practical Applications and Contextual Navigation	8
1.4.3	Research Artifacts	8
1.5	Structure of the thesis	11
2	Towards a Mutable Musical Work: A Literature Review	13
2.1	Evolution of Practice: From Inscription to Generative Systems	13
2.1.1	Material Inscriptions: From Tape to Studio-as-Instrument	14
2.1.2	Formalized Logic: Algorithmic Foundations and Meta-Composition	16
2.1.3	Interactive and Agential Systems	18
2.1.4	Deep Learning in Compositional Practice: The Democratization of Music	19
2.2	Agency, Authorship, and the Theory of Mediation	20
2.2.1	Authorship and Distributed Agency	21
2.2.2	From Artifact to Process: The Ecology of Digital Preservation	22
2.2.3	Technological Agency and the Mediated Work	23
2.2.4	Interfaces as Sites of Co-Creation	24
2.2.5	Autonomous Agents and the Topological Turn	26
2.3	Summary	27
3	From Descriptor to Latent Spaces: A Topographical Review of Musical Models	29
3.1	Descriptor Spaces: The Topography of the Navigable Manifold	29
3.1.1	Tonal Pitch Spaces: The Geometry of Harmony	30
3.1.2	Rhythm and Metric Spaces: Temporal Topology	36
3.1.3	Timbre Spaces: The Multidimensionality of Sound	41
3.2	The Shift to Latent Representations	45
3.2.1	From Prescriptive Features to Emergent Topologies	45
3.2.2	Latent Space Deep Models	46
3.2.3	The Latent Space as a Performative Territory	50
3.3	Corpus-Based Concatenative Synthesis (CBCS) as a Space Enhancer	51
3.3.1	The Topological Framework of CBCS	51
3.3.2	Navigational Strategies in Concatenative Sound Synthesis	52
3.4	Explainable AI in Creative Music Systems	54

3.4.1	Levels of Explainability in Music	54
3.4.2	Latent Disentanglement and Semantic Control	55
3.4.3	Transparency as a Creative Catalyst	55
3.5	Summary: Conceptual and Theoretical Framework of the Mutable Musical Work	56
4	Research Methodology and Design	57
4.1	Problem Statement: Bridging the Gap between Manifold and Performance	57
4.2	Research Questions and Sub-Questions	58
4.2.1	RQ1: On the Ontology of the Mutable Work	58
4.2.2	RQ2: On Practical Applications and Contextual Navigation	59
4.3	Operationalizing the Research Design	59
4.3.1	The Experimental Track: Formalized Metrics and Validation	59
4.3.2	The Practice-Based Track: Iterative Development and Exploration	60
4.3.3	The Iterative Research Cycle	60
4.3.4	The Researcher-as-Navigator: Ethical and Methodological Positioning	61
4.4	Implementation Roadmap: A Two-Phase Study Design	61
4.4.1	Phase One: Symbolic Tonal Representation (Part II)	62
4.4.2	Phase Two: Expressive Timbre Navigation (Part III)	62
4.5	Summary of Research Methodology	63
II	Latent Spaces of Tonal and Structural Dimensions	64
	Introduction to Part II: Latent Spaces of Tonal and Structural Dimensions	65
5	Latent Spaces of Tonal Music	67
5.1	Theoretical Foundations: VAEs and Cognitive Tonal Pitch Space	67
5.1.1	The Geometry of Tonal Pitch Cognition	67
5.1.2	VAEs as Manifold Learners	68
5.1.3	Bridging Latent Embeddings and Semantics	69
5.1.4	The Influence of Music Encoding on Latent Topology	69
5.2	Experimental Design: Comparing Tonal Encodings	70
5.2.1	Methodology	70
5.2.2	Symbolic Music Encodings	71
5.2.3	VAE Model	79
5.3	Evaluation of Latent Space Alignment	83
5.4	Results and Discussion	85
5.4.1	Model Performance and Computational Trade-offs	85
5.4.2	Latent Space Topology and Cognitive Mapping	86
5.5	Summary: Musicological and Creative Implications	87
5.5.1	Analytical Contributions	88
5.5.2	Generative and Creative Potential	88
5.5.3	Summary	89
6	From Fourier Phase Spaces to Latent Representations: Mathematical Structures in VAEs	90
6.1	Defining Theoretical DFT Phase Spaces	91
6.1.1	DFT Phase Space of Pitch Classes	91
6.1.2	DFT Phase Space of Pitch Distributions	92

6.1.3	Comparative Geometry of Pitch and Pitch Class DFT Spaces	92
6.2	Methodology: Correlation between Latent and Phase Spaces	93
6.2.1	Model Architecture and Training	94
6.2.2	Dataset and Augmentation	94
6.2.3	Evaluation Methodology	95
6.3	Results and Discussion	95
6.3.1	Case Study: Chorale <i>BWV 307</i>	96
6.4	Summary: Musicological and Creative Implications	99
7	Sampling Strategies for Musical Generation in Tonal Latent Spaces	100
7.1	The Interpolation Problem in Latent Space	101
7.1.1	Sampling Strategies and Latent Space Topology	101
7.1.2	Continuous Space vs. Discrete Sampling	101
7.2	Comparative Sampling Strategies	102
7.2.1	Linear Interpolation (LI)	103
7.2.2	k -Nearest Neighbors (k -NN)	103
7.2.3	k -NN with Angular Displacement (k -NN+AD)	104
7.3	Evaluation Methodology	106
7.3.1	Materials: Model and Dataset	107
7.3.2	Experimental Hypothesis	107
7.3.3	Computational Method	108
7.3.4	Result Interpretation	108
7.4	Results and Discussion	109
7.4.1	Quantitative Performance Analysis	109
7.4.2	Qualitative Musical Analysis: Case Studies	110
7.4.3	Synthesis: Functional Priority vs. Geometric Pathing	112
7.5	Summary: Implications for Interactive Generative Systems	113
8	BroadcastJSB: Navigating Disentangled Harmonic Spaces	114
8.1	BroadcastJSB: The latent space Radio Receiver	115
8.1.1	The Radio Metaphor as a Cognitive Bridge	115
8.1.2	Dimensionality and Agency	117
8.2	System Architecture	118
8.2.1	Modular Pipeline: From Gesture to Sound	118
8.2.2	Frontend and Interaction (Vue 3 & High-Speed Reactivity)	120
8.2.3	Data Pipeline: From Music21 to JSON Dictionary	124
8.3	Decoder Strategies: Navigating the Manifold	124
8.3.1	NEAR (k -Nearest Neighbors): The Shimmering Manifold	125
8.3.2	MIDPOINT (Linear Interpolation): The Geometric Path	125
8.3.3	VECTOR (Angular Displacement): Tonal Momentum	126
8.4	Evaluation and Discussion	126
8.4.1	Navigation as Analysis	127
8.4.2	The Defamiliarization of the Canon	130
8.4.3	Analog Radio as a Transparency Layer for Visualizing Topological Deviation	130
8.4.4	Limitations and Creative Constraints	131
8.5	Summary: BroadcastJSB as a Demonstration of Co-creative Affordance	132
	Synthesis to Part II: Latent Spaces of Tonal and Structural Dimensions	133

III	Timbre, Gesture, and Motion in Latent Spaces	135
	Introduction to Part III: Timbre, Gesture, and Motion in Latent Spaces	136
9	The Timbre Landscape of the Tenor Saxophone	138
9.1	The RAVE Model and Continuous Timbre Generation	139
9.2	Methodology: Layered Analysis of the RAVE Encoder	140
9.2.1	Pre-trained Models and Training Data	141
9.2.2	SaxTimbre: the Tenor Saxophone Timbre Dataset	141
9.2.3	Evaluation Methodology	144
9.3	Results and Discussion	147
9.3.1	Perceptual Subjective Assessment Results	147
9.3.2	Results of Objective Clustering Analysis	150
9.4	Summary: Creative Affordances and Real-Time Manipulation	152
10	Navigating High-Dimensional Timbre Latent Spaces	155
10.1	The Challenge of Timbre Navigation in Real-Time Systems	156
10.2	Algorithmic Navigation Strategies for Latent Space Exploration	157
10.2.1	k -Nearest Neighbors (k -NN) Binning	158
10.2.2	Distance-Based (BD) Binning	158
10.2.3	Distance with Aligned Trajectory (BD-AT)	159
10.3	Evaluation Methodology and Materials	160
10.3.1	Context and Source Material: <i>Lamento</i>	161
10.3.2	Latent Space Generation: The RAVE Model	162
10.3.3	Generation Parameters and Evaluation Framework	162
10.4	Results and Discussion	163
10.4.1	Perceptual Analysis of Sectional Variations	164
10.4.2	Comparative Analysis of Density and Temporal Continuity	164
10.5	Summary: Navigational Robustness and the Threshold of Coherence	167
11	Mapping Gestural Intent: Taxonomies of Motion in Latent Space	168
11.1	A Taxonomy of Musical Motions	169
11.1.1	The Four Motion Types	169
11.1.2	Consolidating the Taxonomy for the Motiv Dataset	170
11.1.3	Implications for Latent Space Navigation	171
11.2	Experiment: Mapping Geometries of Motion	171
11.2.1	Methodology: Recording Gestural Variations	171
11.2.2	Analysis: Latent Projections and Dimension Weighting	173
11.2.3	Results and Discussion	174
11.2.4	Synthesis: From Geometric Mapping to Gestural Instruction	177
11.3	The Motiv Dataset: Encoding Performance Gesture	178
11.3.1	Recording Protocol and Dataset Construction	179
11.3.2	Setup and Participants	179
11.3.3	Dataset Materials and Guidelines	179
11.3.4	Motion Distribution and Challenges	180
11.3.5	Data Modeling and HDF5 Organization	181
11.4	Preliminary Dataset Analysis	184
11.4.1	Experimental Methodology	185
11.4.2	Results and Discussion	186

11.5 Summary: Mapping Gestural Intent in the Latent Space	192
11.5.1 Summary of Findings	192
11.5.2 Implications for Co-Creative Systems	193
12 Aethra: A Mutable Interface for Mixed Music Navigation	194
12.1 System Overview and Core Architecture	195
12.1.1 Offline Preparation Phase	195
12.1.2 Online Performance System	197
12.2 Case Study: <i>Echoes of the Unseen</i>	207
12.2.1 Material Preparation	208
12.2.2 Timbre Expansion and Real-Time Modulation	208
12.2.3 Improvisational Structure and Formal Score	210
12.2.4 Visual Synthesis and Latent Mapping in TouchDesigner	212
12.2.5 Dissemination and Performance History	213
12.2.6 Creative Affordances of Aethra in Performative Context	214
12.3 Summary: Aethra, A Mutable Interface for Mixed Music Navigation	219
Synthesis to Part III: Timbre, Gesture, and Motion in Latent Spaces	221
IV Towards a Co-Creative Practice: Synthesis and Conclusions	223
Introduction to Part IV: Towards a Co-Creative Practice: Synthesis and Conclusions	224
13 Synthesis and Discussion of Findings	225
13.1 Ontology and Navigation of the Mutable Work (Addressing RQ1)	225
13.1.1 Structural Logic (SQ1.1)	225
13.1.2 Navigational Paradigm (SQ1.2)	228
13.2 On Practical Applications and Contextual Navigation (Addressing RQ2)	229
13.2.1 Material Resistance and Epistemic Action (SQ2.1)	230
13.2.2 Contextual Navigation (SQ2.2)	231
13.3 Summary	233
14 Conclusion	234
14.1 Original Contribution to Knowledge	234
14.2 Limitations and Critical Reflection	236
14.3 Future Work: Toward a Post-Performer Organology	237
14.4 Final Remarks	238
Epilogue: The Manifold and the Muse	239
References	240

List of Figures

3.1	Prototypical Tonnetz representation showing triadic configurations. The highlighted triangle represents a C-minor chord triad, demonstrating how the vertical and horizontal axes organize pitch classes into functional harmonic aggregates.	31
3.2	Spiral Array representation of a minor key as a center of effect (CE). The key is modeled as the CE of its primary triads (i , iv/IV , and V/v), which are themselves generated as the CEs of their constituent pitches on the outermost helix.	33
5.1	Schematic of the modular training pipeline. The training pipeline operates over an EncoderFactory pattern that handles representation-specific logic (e.g., one-hot encoding and decoding) while keeping a consistent ModelTrainer interface. This modularity enables the systematic evaluation of different musical encodings while keeping model hyperparameters and data-handling procedures constant.	71
5.2	Process of encoding and decoding the first measure of <i>Freuet euch, ihr Christen alle Bach (BWV 40/8)</i> as a piano roll (Original encoding, 0-128).	73
5.3	Process of encoding and decoding the first measure of <i>Freuet euch, ihr Christen alle Bach (BWV 40/8)</i> as a piano roll. (Using our continuation-based encoding, 0-128 for attack notes and 129-256 for continuation notes).	74
5.4	Process of encoding and decoding the first measure of <i>Freuet euch, ihr Christen alle Bach (BWV 40/8)</i> with MIDI-like encoding.	75
5.5	Process of encoding and decoding the first measure of <i>Freuet euch, ihr Christen alle Bach (BWV 40/8)</i> with the ABC encoding.	76
5.6	Process of encoding and decoding the first measure of <i>Freuet euch, ihr Christen alle Bach (BWV 40/8)</i> with the Tonnetz encoding. For legibility purposes, the continuations are below the attacks.	78
5.7	Process of encoding and decoding the first measure of <i>Freuet euch, ihr Christen alle Bach (BWV 40/8)</i> with the encoding based on DFT from pitch class distribution (DFT-PC).	80
5.8	Process of encoding and decoding the first measure of <i>Freuet euch, ihr Christen alle Bach (BWV 40/8)</i> with the encoding based on DFT from pitch (DFT-P). For legibility, we only show the output of the first four timesteps.	81

5.9	Schematic of the Implemented Flat Recurrent VAE. The architecture utilizes a symmetric design consisting of a 1024-unit LSTM encoder and decoder to capture long-range temporal dependencies in musical sequences. The latent bottleneck implements the reparameterization trick ($z = \mu + \sigma \odot \epsilon$) to enable gradient-based optimization. The reconstruction loss is dynamically adapted to the input encoding: Mean Squared Error (MSE) for continuous DFT coefficients, Categorical Cross-Entropy (CCE) for token-based MIDI/ABC formats, and Binary Cross-Entropy (BCE) for polyphonic Piano Roll and Tonnetz representations. The model is optimized using the Adam optimizer with a total loss comprising the reconstruction term and a standard KL-divergence ($\beta = 1$).	82
5.10	Distribution of musical (a) modes and (b) keys across the JSB chorale corpus. Blue color indicates minor keys and orange color indicates major keys.	84
	(a) JSB corpus' mode distribution.	84
	(b) JSB corpus' key distribution.	84
5.11	Latent space projections for two chorales using the DFT-P encoding across twelve key augmentations. (a) Displays a perfect positive circular correlation (Kendall's $\tau = 1$). (b) Shows a perfect negative correlation ($\tau = -1$). In both cases, the latent space preserves the circle-of-fifths topology, differing only in the clockwise versus counterclockwise orientation of the key clusters. Cluster coloring and numbering follow the Camelot Wheel convention (keys B through E represented by indices 1–12, with 'B' denoting the major mode). Colored stars indicate the calculated centroids for each key cluster.	88
	(a) Kendall's $\tau = 1$	88
	(b) Kendall's $\tau = -1$	88
6.1	Two-dimensional visualization of major and minor tonalities within the DFT phase spaces of pitch classes ($\{0, 1\}^{12}$) and pitch distributions ($\{0, 1\}^{128}$). Using Locally Linear Embedding (LLE) for dimensionality reduction helps maintain the local symmetries within the space. Importantly, these spaces completely align, reflecting the relationships found in music theory, like the circle of fifths and the connections between major and minor regions.	92
6.2	Two-dimensional LLE visualizations of DFT phase spaces for diatonic chords in C major. The figure depicts DFT $\{0, 1\}^{12}$ (left), DFT $\{0, 1\}^{128}$ (center and right), showcasing sensitivity to octave displacement (center figure, where identical chords are displayed across different octaves) and chord inversions (right figure, where chords are shown in root position and first inversion).	93
6.3	Comparative two-dimensional LLE visualizations of the latent space (left), the $\{0, 1\}^{12}$ pitch class DFT phase space (middle), and the $\{0, 1\}^{128}$ pitch distribution space (right) for the opening phrase of J.S. Bach's chorale <i>BWV 307</i> . Tonal functional clusters are distinguished by color and marker type: Tonic (T) in $\times$, Dominant (D) in $\circ$, and Subdominant (SD) in $\star$. The plots illustrate how the latent space effectively preserves, and in certain instances clarifies, the functional organization found in theoretical common-tone phase spaces.	97

6.4	Analysis of the initial phrase of J.S. Bach’s chorale <i>BWV 307</i> (Bb Major). The musical sequences are organized as follows: (1) the original temporal sequence with functional harmonic analysis; and the chords ordered by proximity to the initial chord (chord 0) within (2) the VAE latent space, (3) the $\{0, 1\}^{12}$ pitch class DFT space, and (4) the $\{0, 1\}^{128}$ pitch distribution space. Red annotations indicate the number of common tones shared with chord 0, while boxed groups denote chords with equivalent distances to the reference.	98
7.1	Sampling procedure for a synthetic chord $\vec{B}'$ using traditional linear interpolation between selected chords A and C. Discrete chords from the corpus are projected into the latent space (blue circles). The geometric midpoint (dark-green hexagon) is calculated as $\vec{z} = 0.5A + 0.5C$. Because this coordinate does not necessarily coincide with an existing musical element, the system samples the nearest available data point ($\vec{B}$) or, to increase variation, the second nearest neighbor ($\vec{B}'$).	104
7.2	Sampling a substitute chord $\vec{B}''$ using k -nearest neighbors (k -NN). Discrete chords from the corpus are projected into the latent space (blue circles). For a given target vector $\vec{B}$, the k nearest neighbors (here, $k = 3$) are identified within the local neighborhood defined by the dotted red boundary. The sampled point $\vec{B}''$ is selected from this verified subset, ensuring the output remains anchored to valid data points while avoiding the low-density regions of the manifold.	105
7.3	Sampling procedure for substitute $\vec{B}'''$ using k -NN+AD relative to the trajectory $[\vec{A}, \vec{B}]$. Discrete corpus chords are represented by blue circles, with the k nearest neighbors of $\vec{B}$ identified within the red dotted boundary. The dark-green arrow denotes the original latent vector $\vec{AB}$. The light-green arrows represent the vectors from $\vec{A}$ to each neighbor candidate. The selection of point $\vec{B}'''$ is attributed to the point whose vector exhibits the minimal angular deviation from the original path, prioritizing directional continuity over raw proximity.	106
7.4	Original sequence, derived from the cadence of the first phrase of J.S. Bach’s Chorale <i>BWV 184.5</i> in D Major, along with five optimal sampling candidates using the three different sampling strategies, ordered by closeness to the original. For each strategy, we present the tonal functions within the phrase, the number of the chord in the dataset of all chords, and the distance metric calculated, i.e., in k -NN and k -NN+AD, we present the Euclidean distance (dist) from the candidate point to the original, plus angular distance for k -NN+AD. For LI, we present the Euclidean distance from the candidate point to the interpolated continuous point in the latent space (dist) and the distance to the original chord (dist to $\vec{B}$).	111
(a)	Original sequence	111
(b)	Linear Interpolation	111
(c)	k -nearest neighbors	111
(d)	k -nearest neighbors and angular displacement	111

7.5	Original sequence, derived from the cadence of the first phrase of J.S. Bach's Chorale <i>BWV 311</i> in B Minor, along with five optimal sampling candidates using the three different sampling strategies, ordered by closeness to the original. For each strategy, we present the tonal functions within the phrase, the number of the chord in the dataset of all chords, and the distance metric calculated, i.e., in k -NN and k -NN+AD, we present the Euclidean distance (dist) from the candidate point to the original, plus angular distance for k -NN with angular displacement (ang). For linear interpolation, we present both the Euclidean distance from the candidate point to the interpolated continuous point in the latent space (dist) and the distance to the original chord (dist to $\vec{B}$).	112
(a)	Original sequence	112
(b)	Linear Interpolation	112
(c)	k -nearest neighbors	112
(d)	k -nearest neighbors and angular displacement	112
8.1	BroadcastJSB's user interface. The design utilizes a mid-century radio metaphor to facilitate the exploration of the Bach chorale latent space. Key interactive elements include: the frequency window (top), which visualizes the drift from the original chords; the station window (center), showcasing the current chorale being played; the tuning dial (bottom center); and the navigation strategy switch (bottom right), which allows users to toggle between NEAR (k -NN), MIDPOINT (LI), and VECTOR (k -NN+AD) sampling strategies.	116
8.2	The bifurcated pipeline architecture of BroadcastJSB. The system decouples the Offline Pre-processing stage (left) from the Real-time Runtime Pipeline (right). During the former stage, the musical corpus is vertically segmented, encoded, and mapped to the MusicVAE latent space. In the runtime environment, user gestural inputs are mapped through a non-linear k -curve to drive an $O(\log N)$ spatial search via a k -d tree. The identified coordinates are then resynthesized in the signal domain using a tuner-modulated additive synthesis engine. This modularity ensures that the interface remains mutable, allowing for the swap of underlying musical corpora without altering the fundamental interaction logic or audio engine.	119
8.3	Frontend project structure of the BroadcastJSB application. The frontend architecture of BroadcastJSB follows a modular logic pattern that explicitly decouples the reactive Vue 3 user interface (<code>App.vue</code> and <code>components/</code>) and static offline datasets (<code>data/</code>) from the underlying computational engines (<code>logic/</code>). This separation ensures that complex state management, spatial searching via k -d trees, and vector mathematics do not compromise the synchronous requirements of the real-time audio synthesis.	121
8.4	Visual feedback mechanisms for latent signal tracking. (a) Latent proximity indicator registering a peak activation state to confirm coordinate alignment with the ground-truth chord (B). (b) Ground-truth indicator exhibiting visual fragmentation artifacts (flicker) during state transitions. Together, these elements translate the calculation of new latent trajectories into legible interface states.	123
(a)	Latent proximity indicator in peak activation state.	123
(b)	Ground-truth indicator with flicker artifacts (on/off).	123

8.5	Comparative harmonic analysis of the initial phrase of J.S. Bach's Chorale <i>BWV 269</i> in G Major. (a) The original sequence, demonstrating a functionally varied common-practice progression. (b) The synthetic transition generated via the MID-POINT (LI) strategy between the phrase's temporal anchors. The generated trajectory illustrates a latent dead zone. Rather than navigating a functionally diverse path, the straight-line mathematical average becomes trapped within a statistically dense subdominant cluster, repeatedly sampling proximal coordinates (e.g., IV_{32}^5 and IV^7) and neutralizing the original cadential tension.	127
	(a) Original.	127
	(b) MIDPOINT.	127
8.6	Harmonic variations of the initial phrase of J.S. Bach's Chorale <i>BWV 269</i> in G Major generated via the NEAR (k -NN) sampling strategy. The subfigures illustrate the effect of a dynamically expanding Euclidean search radius driven by the non-linear tuning input (v). (a) At $v = 0.1$, the system exhibits high structural fidelity, preserving the underlying functional syntax with only minor localized variation (e.g., substituting iii^6 with vi^4). (b) At $v = 0.5$, a moderate spatial expansion introduces further stylistically coherent structural substitutions. (c) At $v = 0.9$, the progression grows substantially more complex, sampling dense, highly chromatic vertical sonorities (e.g., $ii\emptyset^{65b3}$ and IV^{6b5}) that strictly belong to the established corpus. (d) At $v = 1.0$, stylistic integrity completely breaks down; structural coherence dissolves, and the final cadential resolution is neutralized by functionally ambiguous, high-entropy aggregates such as $biii_{42}^7$	128
	(a) NEAR $v = 0.1$	128
	(b) NEAR $v = 0.5$	128
	(c) NEAR $v = 0.9$	128
	(d) NEAR $v = 1.0$	128
8.7	Harmonic variations of the initial phrase of J.S. Bach's Chorale <i>BWV 269</i> in G Major generated via the VECTOR (k -NN+AD) sampling strategy. The subfigures demonstrate the effect of preserving directional tonal momentum as the tuning input (v) increases, driving the sequence along specific geometric trajectories rather than maintaining localized proximity. (a) At $v = 0.1$, the progression exhibits immediate structural divergence, establishing a new functional path (e.g., shifting from I to VI by the second chord). (b) At $v = 0.5$, the sustained angular displacement propels the harmony into denser, highly chromatic regions of the manifold, generating intentional momentum sonorities such as bII_{53}^{6a} and $vii\emptyset_{43}^{65}$. (c) At $v = 0.9$ and (d) $v = 1.0$, the continuous geometric extrapolation forces the sequence into a specific mathematical "sink" or functionally ambiguous "interstation noise" region. Across high-deviation thresholds, the trajectory reliably converges on the high-entropy aggregate $\sharp vi_{42}^{ob6}$, fully dissolving the traditional boundaries of the established diatonic grammar.	129
	(a) VECTOR $v = 0.1$	129
	(b) VECTOR $v = 0.5$	129
	(c) VECTOR $v = 0.9$	129
	(d) VECTOR $v = 1.0$	129
9.1	Comparison of the input spectrogram for a soft low C4 recording against the reconstructed outputs from latent dimensions 1, 11, and 16 of the VocalSet latent space. The visualization highlights the progressive transformation of spectral content and the specific low-register reactivity observed in layer 16.	148

(a)	Input Audio	148
(b)	Latent Dimension 1	148
(c)	Latent Dimension 11	148
(d)	Latent Dimension 16	148
9.2	Comparative spectrograms of the input audio for multiphonic aggregate no. 5 alongside reconstructions decoded from latent dimensions 1, 11, and 16 of the VocalSet latent space. The visualization demonstrates the varying degrees of spectral fidelity across the model hierarchy and the specific challenges of representing non-harmonic partials in deeper encoder latent dimensions.	149
(a)	Input Audio	149
(b)	Latent dimension 1	149
(c)	Latent dimension 11	149
(d)	Latent dimension 16	149
9.3	Davies-Bouldin Score Results by latent dimension. The best results are marked with a circle. All audio is normalized for the cluster categories corresponding to extended techniques and pitch registers.	151
(a)	Clustering on Technique	151
(b)	Clustering on Register	151
(c)	Clustering on Dynamic	151
(d)	Combined Clustering	151
10.1	Sampling of an audio unit within a two-dimensional t-SNE projection of the RAVE latent space using the k -NN binning strategy. Discrete latent units are represented as points, with the reference unit (A) indicated by a blue arrow. The color gradient signifies the organization of units into discrete “bins” based on spatial proximity, where each bin corresponds to a specific value of the closeness parameter C	159
10.2	Sampling of an audio unit within a two-dimensional t-SNE projection of the RAVE latent space using the BD binning strategy. Discrete latent units are represented as points, with the reference unit (A) indicated by a blue arrow. The color gradient signifies the organization of units into discrete bins based on spatial proximity, where each bin corresponds to a specific value of the closeness parameter C	160
10.3	Sampling of an audio unit within a two-dimensional t-SNE projection of the RAVE latent space using the BD-AT binning strategy. Discrete latent units are represented as points, with the reference unit (A) indicated by a blue arrow. The color gradient signifies the organization of units into discrete bins based on spatial proximity, where each bin corresponds to a specific value of the closeness parameter C	161
11.1	Illustration of the four motions within Pinho’s taxonomy (Pinho, 2016). Left-to-right flow indicates the motion from pre-existing musical material to the present manifestation of musical discourse.	170
11.2	Musical score of phrase 1 from the saxophone part of Villa-Rojo (1989)’s <i>Lamento</i> , corresponding to 0:18–0:38.	172
11.3	Musical score of phrase 2 from the saxophone part of Villa-Rojo (1989)’s <i>Lamento</i> , corresponding to 3:30–3:48.	172
11.4	Musical score of phrase 2 from the saxophone part of Villa-Rojo (1989)’s <i>Lamento</i> , corresponding to 7:30–7:45.	173

11.5	Two-dimensional t-SNE projection of RAVE latent trajectories (units 50–60) for the second saxophone phrase. The visualization illustrates the spatial dilation of contrary and oblique motions compared to the overlapping pathways of parallel transpositions. Motion types are color-coded with start points (circles) and end-points (wedges) indicating temporal direction within the manifold.	174
11.6	Latent space trajectories (dims 0, 5, 10, 15) for saxophone phrase 2. The comparison highlights structural consistency in most dimensions, alongside pitch-sensitive variance in dimension 10. Directional markers (circles to wedges) and color-coding track temporal flow and motion types across units 50–60.	175
(a)	Dim. 0	175
(b)	Dim. 5	175
(c)	Dim. 10	175
(d)	Dim. 15	175
11.7	Global t-SNE projection of latent trajectories (units 50–60) for the first electronic phrase. The visualization highlights the significant spatial isolation of the original electronic source from all saxophone-based variations and mimics. Parallel, oblique, and contrary motions are color-coded to illustrate the varying degrees of spatial dilation and directional alignment relative to the saxophone mimic.	176
11.8	Individual latent trajectories for dimensions 0, 5, 10, and 15 (units 50–60) of the first electronic phrase. The plots reveal a distinct spatial offset in dimension 10, indicating the latent sensitivity to recording density and tape mastering versus live saxophone replication. In dimensions 5 and 15, parallel motions demonstrate high directional correlation with the electronic source. In contrast, contrary motions exhibit significant divergence in the technical register (dim. 15). Temporal sequence is indicated by filled circles (start) and upward wedges (end).	177
(a)	Dim. 0	177
(b)	Dim. 5	177
(c)	Dim. 10	177
(d)	Dim. 15	177
11.9	Examples of parallel motions applied to the saxophone score of phrase 1.	180
(a)	Transposition by a downward minor third.	180
(b)	Dynamics and articulation variations.	180
11.10	Examples of oblique motions applied to the saxophone score of phrase 1.	181
(a)	Dynamics and interval inversion on appoggiaturas.	181
(b)	Timbre changes with added trills and larger intervals in the appoggiaturas.	181
11.11	Examples of contrary motions applied to the saxophone score of phrase 1.	181
(a)	Lower pitch register, timbre, and dynamic changes using subtones.	181
(b)	Higher pitch register and timbre changes utilizing trill additions.	181
11.12	Latent vectors corresponding to a phrase lasting one second result in 23 compressed audio frames, leading to a latent matrix that is a 23 by 16-dimensional vector containing floating-point numbers. Dimensions in the figure are labeled d1 to d16, with timesteps ranging from 0 to 22.	183
11.13	Structural organization of the Motiv dataset.	184
11.14	Visualization of high-dimensional trajectories via t-SNE, comparing an original phrase with its most structurally similar parallel variation (PCC = 0.8155). The grouping illustrates how non-transpositional parallel elaborations maintain cohesive patterns in the latent space. Frame windowing was increased to 12 frames to reduce noise in the two-dimensional representation.	187

11.15	Average PCC values for taxonomic motion types (P, O, C) per saxophonist. Values indicate the degree of structural resemblance between recorded variations and the source phrases within the RAVE latent space, with higher values denoting greater structural alignment.	188
11.16	Angular distance values between centroids for different motion types (P, O, C) across four saxophonists are analyzed. Lower angular distances indicate a stronger similarity between the recorded motion's directionalities and the original musical phrase.	189
11.17	Cosine similarity values for different motion types (P, O, C) across four saxophonists are analyzed. Higher cosine similarity indicates a stronger similarity between the recorded motion's directionalities and the original musical phrase.	190
11.18	Mean Euclidean distances for each player-motion combination. This visualization highlights the variability in deviation from the original phrase across different motion types and performers.	192
12.1	Architectural Overview of the Aethra System. The workflow is bifurcated into two distinct operational stages: (left) the offline preparation phase, utilizing a Google Colab notebook to encode the fixed audio tape into serialized latent representations; (right) the online performance phase, operating within PD via the py4pd library. The real-time environment facilitates navigation of the preloaded latent space via a single-axis control knob. Additionally, the architecture incorporates a dual-mode navigation engine, dynamic latent dimension weighting, real-time granular audio synthesis, and an OSC data stream to a TouchDesigner visual engine for real-time topological rendering.	196
12.2	Main Performance Interface of the Aethra System in PD. It centralizes control of the navigation in a single knob interface. The user can also toggle between the two different navigation strategies (i.e., HMM/ k -NN), between synthesis logic modes (i.e., linear fidelity/cumulative transformation), and set latent dimension weights for proximity calculation.	198
12.3	PD Interface for Asset Initialization and Global Configuration. The interface displays the management of core system dependencies, including the Latent Space Loader and the audio tables. The initialization relies on a configuration parser that automates the retrieval of file paths for the audio source, saved latent vectors (.pkl files), and HMM transition models. This way, everything can work smoothly together between the Python backend and the real-time synthesis engine.	199
12.4	One-knob interface C during distance-based exploration. The ticks in the knob demonstrate that C works as a continuous value, mapping the input to latent similarity bins.	201
12.5	One-knob interface C during motion-based Mode. The ticks in the knob show how it acts as a discrete selector for original, parallel, oblique, and contrary HMM models.	202
12.6	Architecture of the navigation logic and synthesis control. The patch integrates the <code>py.navigation_mechanism</code> to facilitate real-time latent space selection. The system distinguishes between navigation modes, motion-based and distance-based (<code>\$0-cl-mode</code>), and synthesis logic, where <code>\$0-mode</code> toggles between Cumulative (generative sequencing) and Linear (score-following) synthesis logic.	204

12.7	Segment indexing and buffer addressing logic. The integer segment index (<i>segNr-play.r</i>) received navigation layer is multiplied by a constant unit size of 8192 samples to determine the precise start-read position within the <i>\$0-sample</i> audio buffer. The unit is retrieved and played with a cloned granular playback abstraction (<i>abs/player.n</i>). A Hann windowing strategy (<i>\$0-hann</i>) is employed to ensure morphological continuity during real-time navigation of the latent space. .	205
12.8	Schematic of the real-time granular playback engine. The 170ms metro cycle controls how often the granular envelopes activate, creating a smooth and evolving sound. This regular triggering helps blend different segments from the latent space into a unified sound while keeping a consistent shape as the synthesis events flow continuously.	206
12.9	Interface (a) and implementation (b) of the real-time timbre transfer module in PD using the <i>nn_tilde</i> object with the MusicNet model. The architecture illustrates the encoder and decoder stages of the process, with a 16-channel set of sliders used to modulate latent dimension weights.	209
	(a) Interface of real-time timbre transfer module.	209
	(b) Implementation of real-time timbre transfer module.	209
12.10	Representative excerpt of the <i>Echoes of the Unseen</i> auto-score. The notation facilitates a state of contextual navigation by providing a formal skeleton for the performer to explore the high-dimensional latent space, mediating the vulnerability of free improvisation through pre-defined timbre landmarks and extended techniques.	211
12.11	<i>Echoes of the Unseen</i> 's visual engine in TouchDesigner. The interface displays a point-cloud representation of the latent space, a proximity cursor for real-time navigation, and a morphological feedback slider in the background, as well as the architectural components that comprise the visual system in the foreground. . . .	212

List of Tables

3.1	Typology of Tonal and Pitch Space Models in Music Theory and Computation. . .	31
3.2	Typology of Rhythm and Metric Space Models: From Symbolic Hierarchies to Continuous Landscapes.	37
3.3	Typology of Timbre Space and Audio Descriptor Models.	42
3.4	Correlation between acoustic descriptors and timbre interpretation.	43
3.5	Comparison Between Traditional Descriptor Spaces and Neural Latent Spaces. . .	50
5.1	Quantitative results on the VAE model’s reconstruction performance across the six presented encoding schemes: piano roll (P), MIDI-like (M), ABC (A), Tonnetz (T), DFT of pitch classes (DFT-PC), and pitch (DFT-P) distributions. ABC and MIDI-like encodings yielded the highest reconstruction accuracy and lowest KL-divergence.	86
5.2	Average and standard deviation values for the cluster and key-distance metrics, assessing alignment with cognitive pitch spaces for the six presented encoding schemes: piano roll (P), MIDI-like (M), ABC (A), Tonnetz (T), DFT of pitch classes (DFT-PC), and pitch (DFT-P) distributions. DFT-P outperforms other encodings in all metrics.	86
6.1	Pearson r and Spearman ρ correlation values for the corpus’s latent spaces and DFT phase distances.	96
7.1	Results of the comparative analysis regarding the perceptual relatedness of the chord sampling strategies across major, minor, and combined datasets.	109
9.1	Pre-trained RAVE models employed in our experiments, including their architectural versions, sample rates, latent dimensionality, and training dataset descriptions.	142
9.2	Collection of sounds encompassing the entire spectrum of the tenor saxophone (timbre space).	143
9.3	Comparative analysis of clustering performance across all RAVE model latent dimensions. Optimal values for each categorical framework and metric are indicated in bold. The labels N and NN distinguish between evaluations performed with normalized and non-normalized audio input, respectively, highlighting the impact of signal amplitude on latent organization.	150
9.4	Qualitative and quantitative summary of RAVE model reactivity to tenor saxophone timbre. The table synthesizes expert perceptual assessments with mathematical clustering metrics (Davies-Bouldin and Dunn Indices) to define the creative transparency and structural limitations of each architecture.	153

10.1	Mean similarity metrics for latent space navigation strategies. Higher values indicate greater preservation of original musical features (zcr, rhythm, chroma, energy, spectral contrast, and perceptual) relative to the source corpus.	165
11.1	Percentage of musical samples exhibiting each transformation category within the Motiv dataset. The data illustrate the dominance of direct transposition in parallel motions (55%) and the prevalence of phrase reversals and dynamic inversions in contrary motions (50% and 30% respectively).	182
11.2	Pearson Correlation Coefficients (PCC) between recorded motion types and original phrases across four saxophonists. Unless otherwise noted with * ($p < 0.05$) or † (not significant, $p > 0.05$), all values indicate $p < 0.0001$. Higher values reflect stronger structural similarity in the latent space.	186
11.3	Angular distances between the centroids of recorded motion trajectories and their respective original phrases. Lower values indicate a stronger directional alignment within the latent space, while higher values signify greater expressive divergence from the source material.	188
11.4	Cosine similarity between synchronized latent trajectories of recorded motion types and original phrases. Values closer to 1 indicate stronger directional alignment, whereas values near zero or negative indicate structural divergence.	190
11.5	Euclidean distances between latent trajectories of recorded motion types and original phrases. Higher values represent a greater absolute spatial deviation (variation) from the original performance manifold.	191
12.1	Comparison of Navigation Strategies within Aethra-PD.	200
12.2	Latent Dimension Weight Configurations ($z_1 - z_{16}$) for Performative Presets.	210
12.3	Performance and Dissemination History of <i>Echoes of the Unseen</i>	214
13.1	Synthesis of the empirical and performative findings of this research, organized by the two primary research pillars and the dual domains of Tonal and Timbre logic.	226

Abreviaturas e Símbolos

AI	Artificial Intelligence
ANN	Artificial Neural Network
ANT	Actor-Network Theory
ATN	Augmented Transition Network
CAC	Computer-Assisted Composition
CAAC	Computer-Aided Algorithmic Composition
CBCS	Corpus-Based Concatenative Synthesis
CVAE	Conditional Variational Autoencoder
DL	Deep Learning
DFT	Discrete Fourier Transform
GUI	Graphical User Interface
IIL	Intelligent Instruments Lab
IML	Interactive Machine Learning
IMS	Interactive Music Systems
IRCAM	Institute for Research and Coordination in Acoustics/Music
KL	Kullback–Leibler (Divergence)
k-NN	k-Nearest Neighbors
k-NN+AD	k-Nearest Neighbors with Angular Displacement
LI	Linear Interpolation
LLM	Large Language Model
LSTM	Long Short-Term Memory
MDS	Multidimensional Scaling
MFCC	Mel-Frequency Cepstral Coefficients
MIDI	Musical Instrument Digital Interface
MIR	Music Information Retrieval
ML	Machine Learning
MSE	Mean Square Error
PCA	Principal Component Analysis
PD	Pure Data
RIM	Réalisateur en Informatique Musicale
RNN	Recurrent Neural Network
SOM	Self-Organizing Map
STFT	Short-Time Fourier Transform
t-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
VAE	Variational AutoEncoder
RAVE	Realtime Audio Variational AutoEncoder
XAI	Explainable AI

Part I

Framing the Mutable Musical Work

Chapter 1

Introduction

This thesis explores the transformation of musical works in the age of advanced computational technology. Traditionally, Western art music has embraced the idea of the work-concept, which views a musical composition as a stable, written entity that retains its identity through accurate notation and faithful performances (Goehr, 1994). However, the rise of electroacoustic practice and interactive digital systems has progressively challenged this hierarchy, demanding new frameworks that account for the fluid boundaries between composition, performance, and technological agency.

This introductory chapter establishes the conceptual foundation for the mutable musical work, proposing a shift from the static score to the dynamic latent space.¹ Rather than viewing machine learning (ML)² just as a manner of automating tasks, this research presents it as a site of computational openness, where the musical work's identity is maintained not just via a fixed sequence of notes, but through a navigable geometry of possibilities. By conceptualizing performance as an exploration of high-dimensional state spaces, we move toward a model of contextual navigation, where the machine's encoded contextual knowledge and the performer's gestural intent engage in a closed-loop dialogue of structural discovery.

To establish the theoretical framework for the mutable musical work, this chapter traces the evolution of the musical object from a static artifact to a dynamic process through several key thematic moments. Section 1.1 begins by examining the historical tension between traditional notation and machine learning, identifying the need for new ontological definitions that account for the fluid boundaries of contemporary digital practice. Building on this need, Section 1.2 proposes re-framing the musical work as a topological manifold, in which structural logic and performative potential are unified within high-dimensional latent spaces.

¹ A latent space is a low-dimensional, continuous, and often curved structure that underlies high-dimensional data, representing the essential, hidden, or latent features of that data (Kingma and Welling, 2014).

² Machine learning is a subfield of artificial intelligence characterized by the development of statistical algorithms that enable computational systems to autonomously induce patterns, structures, or predictive models from large-scale datasets without being explicitly programmed for every contingency. Unlike classical, rule-based programming, which relies on a deterministic, if-then, logic, ML models utilize iterative optimization processes to minimize a loss function, effectively transforming the computer into a learner that can generalize from observed data to novel, unseen scenarios (Michalski et al., 1983).

This shift in representation necessitates a new performative model, which Section 1.3 addresses by analyzing the redistribution of agency in co-creative music systems, ultimately defining the performer's role as a navigator within a contextually bounded environment.

Following this theoretical development, Section 1.4 outlines the thesis's primary objectives (i.e., the challenge of making black box AI intelligible) and details the practical contributions this research offers to musicology and computational creativity. Finally, Section 1.5 provides a roadmap for the subsequent parts of the thesis, tracing the progression from the symbolic tonal explorations of Part II to the sub-symbolic timbre landscapes investigated in Part III.

1.1 The Problem of the Fixed Work in Digital Practice

In contemporary electroacoustic music, the adoption of recorded media fundamentally altered the relationship between the composer's intent and the performer's interpretation. This is most evident in the tradition of mixed media (or *musique mixte*), where acoustic instruments are paired with fixed electronic counterparts, typically stored as pre-recorded audio tape (Manning, 2013). While this approach expands the available sonic resources, it creates a significant performative challenge: the fixed nature of the electroacoustic part imposes strict metronomic constraints, restricting the performer's rhythmic flexibility and expressive timing to the pre-determined timeline of the machine (Emmerson, 2007).

A central issue in contemporary digital practice is the established asymmetry of agency between human and machine. Within standard Computer-Assisted Composition (CAC) and traditional live-electronics environments, historically exemplified by the research models at centers like IRCAM (Institute for Research and Coordination in Acoustics/Music), the computer is generally structured as a musical assistant or a computational processor that facilitates the translation of musical concepts into scientific data (Zattra, 2018a; Boulez and Gerzso, 1988). In this paradigm, limiting the computer's autonomy is a deliberate design choice rooted in the Western classical tradition of maintaining strict compositional control. Technologies such as advanced score followers (e.g., Antescofo) are designed to consolidate the composer's control, reinforcing the central role of the written score and utilizing the machine as a direct executor of its instructions (Tiffon, 2018; Cont, 2008). Consequently, unexpected generative outputs from the machine are typically classified as system errors rather than expressions of machine agency, as they represent a breach in the expected control parameters (Nika, 2016; Fels, 2004).

To preserve this control hierarchy, traditional electronic systems are often designed for psychophysiological transparency, prioritizing a direct, predictable mapping between physical control gestures and acoustic output (Fels et al., 2002; Miranda, 2014). However, by emphasizing effortless and transparent one-to-one interactions, these architectures effectively relegate the system to a passive tool rather than a collaborative partner (Fels et al., 2002; Ryan, 1991). As Ryan (1991) observes, designing interfaces solely to minimize physical effort can inadvertently widen the cognitive separation between the musician and the underlying generative processes. In contrast, for a computational framework to function as a genuine performative agent, it necessitates a degree

of designed unpredictability and material resistance (Ryan, 1991). Consequently, incorporating an ungainly or nonlinear component deliberately fractures the direct continuity between physical gesture and the resulting sound, compelling the performer to actively adapt to and negotiate with the system in real time (Granzow, 2018).

Furthermore, the lack of a performative role for the computer results from agency being displaced into the software's architecture prior to the performance (Rowe, 1999). In many mixed-media works, the generative capacity of the computer is fully specified during the composition phase, reducing the machine to the execution of fixed operations during live performance (Garnett, 2001; Blackwell et al., 2012). Consequently, the system lacks the capacity for reciprocal interaction, or what George Lewis terms a "dialogic imagination" (Lewis, 2000). Lewis provides a prominent counter-model with his system *Voyager*, asserting that a genuinely interactive system should function as a "virtual improvising orchestra" capable of generating complex responses to human input alongside independent behaviors driven by internal processes (Lewis, 2000; Born, 2005). Without the ability to produce autonomous, unscripted outputs that challenge the human musician, which Chadabe (1984) characterizes as responding in a "complex, not entirely predictable way," the computer functions merely as a precise timing device or signal processor, lacking the autonomy necessary to serve as an adaptive, collaborative partner.

Consequently, when computational systems are restricted to predetermined execution, the musical work is reduced to a static artifact (Jones et al., 2012). This structural rigidity highlights a fundamental conflict in digital music practice: balancing the autonomous variation afforded by generative systems with the composer's need for precise expressive control (Rowe, 1999; Jones et al., 2012; Collins, 2008). As the musical work integrates high-dimensional machine learning models, its ontology shifts significantly. Rather than functioning as a linear set of symbolic instructions to be interpreted, the composition becomes a multi-dimensional state space of possibilities (Magnusson, 2019; McDermott, 2021).

Existing frameworks frequently struggle to accommodate this algorithmic flexibility without compromising the structural intent of the human creator (Nika, 2016; Jones et al., 2012). Addressing this limitation requires a departure from both the fixed electronic track and the paradigm of the computer as a passive assistant, moving instead toward the concept of the mutable interface (Emmerson, 2007; Drummond, 2009). Within this framework, the musical work is no longer treated as a finalized product, but rather as a navigable territory that facilitates an active, structural negotiation of agency between the human and the computational system (Magnusson, 2019; Di Scipio, 2003; Lewis, 2000).

1.2 The Mutable Interface as a Research Framework

The transition from the open work to the mutable interface represents an ontological shift in the conceptualization of musical structure. While Eco (1989) defines openness as a field of interpretive possibilities structured by the composer, contemporary digital systems introduce a computational

openness in which the structural field itself is dynamic (Magnusson, 2019). Framing the musical work as a topological manifold advances beyond the symbolic indeterminacy characteristic of twentieth-century aleatoric music. Whereas aleatoric works rely on discrete structural permutations, the latent space of a deep generative model functions as a continuous, multi-dimensional representational space (McDermott, 2021; Yee-King, 2022). In this context, mutability is not a secondary byproduct of performance interpretation, but a fundamental property of the system's digital encoding. This paradigm transforms musical performance from the execution of a predetermined plan into the active navigation of a latent space (Yee-King, 2022; Magnusson, 2019).

Building upon this premise of shared discovery, the nature of musical interaction requires a fundamental reconsideration. It is no longer sufficient to view the computer merely as a reactive mirror. Instead, the relationship shifts toward an ecosystemic model. As Di Scipio (2003) argues, the composer's task in this paradigm is no longer to organize a linear sequence of sound objects, but to creatively "compose the interactions" themselves. Within this audible ecosystem, the human performer, the computational system, and the physical environment are coupled in continuous feedback loops, constantly negotiating their roles and mutual limits (Waters, 2007; McCormack, 2012).

This structural coupling challenges the traditional boundary between the human subject and the technological object. Drawing upon the philosophy of agential realism, recent research into dynamic and agential scores suggests that these performative relationships are most accurately understood through the concept of intra-action (Armitage and Magnusson, 2023; Armitage et al., 2024a; Barad, 2007). Unlike standard interaction, which assumes that pre-existing, separate entities exchange data, intra-action posits that the agencies of the performer and the algorithmic system do not precede their encounter. Instead, they are mutually constituted through their active entanglement during the performance (Barad, 2007; Armitage and Magnusson, 2023). The performance emerges not from a human commanding a passive machine, but from a continuous exchange of "mutual provocations" or "intra-active perturbations" (Armitage et al., 2024a).

To effectively sustain this intra-active engagement, the computational system must exhibit a requisite degree of unpredictability and structural complexity. If the system's responses remain entirely deterministic and perfectly transparent, it inevitably reverts to the status of a passive tool, failing to provide the material resistance necessary for a genuine musical dialogue (Chadabe, 1984; Ryan, 1991). Consequently, frameworks such as Live algorithms (Blackwell et al., 2012) stipulate that an artificial musical partner must demonstrate not only participation and novelty but also "leadership." By generating encoded structural responses and occasionally providing topographical pushback to guide the performer's exploration, the system successfully establishes a non-hierarchical, entangled performative network (Lewis, 2000; Born, 2005).

Ultimately, the mutable interface redefines the act of composition and the ontology of the musical work. The composer designs the latent space of the system, establishing a boundary object that facilitates the emergence of distributed agency (Armitage and Magnusson, 2023; Jones et al., 2012). The musical work, therefore, ceases to be a static artifact and exists instead as a set of dynamic potentials that are actualized only through the situated, real-time negotiation between the

embodied human performer and the intelligent computational environment (Vear, 2019; Xambó and Roma, 2024).

By shifting the focus from the interpretation of a static score to the navigation of a dynamic state space, a process aligned with Magnusson (2019)’s concept of *ergodynamics*,³ this research explores how computational technology facilitates a structural openness that redefines the traditional roles of composer and performer (Eco, 1989; Jones et al., 2012). The systems developed in this thesis, BroadcastJSB and Aethra, are situated within this trajectory. Rather than functioning as passive tools for sound production, they operate as mutable environments where authorship is distributed through continuous, real-time interaction (Jones et al., 2012; Rutz, 2016). This shift necessitates a framework of contextual navigation, wherein the machine’s encoded material resistance establishes a continuous feedback loop (Ryan, 1991; Waters, 2007; Di Scipio, 2003). Consequently, the performance emerges from the dynamic interplay between human intentionality and the system’s defined topological constraints (Dahlstedt, 2009).

1.3 Contextual Navigation in Machine-Mediated Performance

Building on the concepts of distributed agency and technological actants, this research advances a performative framework of contextual navigation (Latour, 2005; Stapleton and Davis, 2021). In traditional musical performance, intentionality is typically viewed as a human-centric attribute, wherein the performer directs the execution toward a specific musical goal (Ingarden et al., 1986; Anscombe, 2020). However, in interactive digital systems, the interaction shifts from a unidirectional command structure to what Vear (2019) conceptualizes as a dynamic computational environment. Rather than attributing actual intentionality or an independent mind to the machine, the system functions through simulated agency, operating as a highly context-aware framework that the performer navigates (Winkler, 2001; Hatten, 2018).

The machine ceases to function merely as a passive tool. Instead, it leverages the statistical structures and learned manifolds of its training data to establish a probabilistic mapping of the musical domain (McDermott, 2021; Yee-King, 2022). When a performer interacts with a latent space, such as the timbre manifold of the RAVE model (Caillon and Esling, 2021), the system does not strictly mirror the input (Tatar et al., 2021; Blackwell et al., 2012). Rather than generating entirely autonomous outputs, the machine produces contextually bounded possibilities derived from its internal probability topology (Briot et al., 2020; McDermott, 2021). This encoded structural knowledge actively restricts the generative space, providing the necessary material resistance and unpredictability for interaction (Ryan, 1991; Chadabe, 1984). Consequently, the performer is required to actively adapt and negotiate with the system in real time, establishing a continuous feedback loop where the system’s topological boundaries influence the human’s subsequent gestures, and vice versa (Di Scipio, 2003; Xambó and Roma, 2024).

³Magnusson (2019) defines *ergodynamics* (from the Greek *ergon* for work and *dynamis* for potential or power) as a musical instrument’s inherent capacity for expression and discoverability. Distinct from simple physical affordances and constraints, the concept functions as a comprehensive framework for evaluating the material, historical, and operational qualities that determine how an instrument is actively explored, learned, and integrated into musical practice.

Crucially, [Vear \(2019\)](#) observes that such systems can facilitate a state of flow, effectively blurring the boundaries between musician, instrument, and score ([Vear, 2019](#); [Nijs et al., 2013](#)). In this state, the performance manifests as an emergent property of the coupled system, situated neither exclusively in human cognition nor entirely within the machine’s code, but within the active interaction between the two ([Kim and Seifert, 2007](#); [DiBlasi, 2022](#)). The necessity for navigation arises directly from this encoded material resistance ([Ryan, 1992](#)). Because the system’s topology is embedded in a high-dimensional state space that precludes linear, one-to-one control, the performer must function as a navigator within a bounded representational space ([McDermott, 2021](#); [Wessel and Wright, 2002](#); [Jones et al., 2012](#)). This approach aligns with what [Paine \(2002\)](#) describes as “dynamic coupling,” wherein the identity of the musical work emerges from the ongoing interplay between human intentionality and the system’s defined topological constraints ([Blackwell et al., 2012](#); [Dahlstedt, 2009](#)).

1.4 Research Objectives, Questions, and Contributions

To investigate the transition of the musical work from a fixed artifact to a mutable interface, this thesis addresses two primary research questions. Each question is paired with specific objectives and results in distinct peer-reviewed or technical contributions.

1.4.1 RQ1: On the Ontology of the Mutable Work

The first research question, RQ1, investigates the ontology of the mutable work by asking:

Can high-dimensional latent spaces be structured to reflect perceptual, mathematical, and music-theoretical concepts, and what navigational paradigms allow these manifolds to function as intelligible, mutable musical works?

To address this question, this research pursues two primary objectives. The first one, objective 1.1, seeks to formalize a mutable musical work as a structured latent space where musical identity is defined by geometric continuity and aligned with formal logic. Furthermore, objective 1.2 focuses on developing a navigation-based interaction model that replaces traditional score-following with real-time latent space exploration.

These objectives are realized, at the tonal level, through the theoretical frameworks explored in “Exploring Latent Spaces of Tonal Music using Variational Autoencoders” ([Carvalho and Bernardes, 2023](#)) and “Fourier (Common-Tone) Phase Spaces are in Tune with Variational Autoencoders’ Latent Space” ([Carvalho and Bernardes, 2024b](#)), and the navigation strategies proposed in “Exploring Sampling Strategies in Latent Spaces for Music Generation” ([Carvalho and Bernardes, 2024a](#)). At the timbre level, they are supported with the creation of the SaxTimbre ([Carvalho et al., 2024](#); [Carvalho and Sousa, 2024](#)) and Motiv ([Carvalho et al., 2025a](#); [Carvalho and Sousa, 2025b](#)) datasets, and the study “Exploring Timbre Latent Spaces: Motion-Enhanced Sampling for Musical Co-Improvisation” ([Carvalho et al., 2025b](#)). These datasets and navigation strategies provide the empirical ground truth for timbre latent representation and gestural navigation.

1.4.2 RQ2: On Practical Applications and Contextual Navigation

The second research question investigates the performative dynamics and the redistribution of agency within the interface:

In what ways does the evolution of the epistemic tool (Magnusson, 2009) redistribute authorship, and how can contextual navigation be fostered through the machine’s material resistance?

To address this question, this research pursues two primary objectives. The first, objective 2.1, suggests designing an interactive loop where the machine’s contextually bounded possibilities and the high-dimensional entanglement of the manifold provide material resistance to the performer’s gestural intent. The second, objective 2.2, evaluates this contextual navigation through performative practice, shifting the performer’s role from controller to navigator. This investigation is documented in “Exploring Timbre Latent Spaces: Motion-Enhanced Sampling for Musical Co-Improvisation” (Carvalho et al., 2025b) and the development of the Aethra system (Carvalho, 2026b).

To address this question, objective 2.1 suggests designing an interactive loop where the machine’s contextually bounded possibilities and the high-dimensional entanglement of the manifold provide the material resistance necessary to transform the system into a dynamic, mutable interface. Objective 2.2 evaluates this contextual navigation through performative practice, shifting the performer’s role from controller to navigator. These investigations are implemented and documented through the BroadcastJSB web-application (Carvalho, 2026a), and the Aethra performance system (Carvalho, 2026b).

1.4.3 Research Artifacts

To provide a transparent view of the technical outputs, the practical contributions are categorized into publications, software systems, and research datasets. Each repository represents a specific implementation of the mutable interface framework, ranging from symbolic tonal navigation to sub-symbolic timbre exploration. A complete and updated collection of these resources can be accessed through a curated website.⁴

1.4.3.1 Publications

To deliver a clear overview of the research outputs, the following list summarizes the peer-reviewed publications and conference proceedings that constitute the core of this thesis.

- Carvalho, N., & Bernardes, G. (2023). Exploring Latent Spaces of Tonal Music using Variational Autoencoders. The International Conference on AI and Musical Creativity (AIMC). DOI: 2311.03621

⁴<https://nadiacarvalho.github.io/MutableMusicalInterface>, last accessed January 29, 2026.

- Carvalho, N., & Bernardes, G. (2024). Exploring Sampling Strategies in Latent Spaces for Music Generation. *Proceedings of the 21st Sound and Music Computing Conference (SMC)*, 74–81. DOI: [10.5281/zenodo.14336029](https://doi.org/10.5281/zenodo.14336029)
- Carvalho, N., & Bernardes, G. (2024). Fourier (Common-Tone) Phase Spaces are in Tune with Variational Autoencoders' Latent Space. In T. Noll, M. Montiel, F. Gómez, O. C. Hamido, J. L. Besada, & J. O. Martins (Eds.), *Mathematics and Computation in Music: 9th International Conference, MCM 2024, Coimbra, PT, June 18–21, Proceedings. Lecture Notes in Computer Science* (Vol. 14639, pp. 305–316). Springer, Cham. DOI: [10.1007/978-3-031-60638-0_24](https://doi.org/10.1007/978-3-031-60638-0_24)
- Carvalho, N., Sousa, J., Bernardes, G., & Portovedo, H. (2024). Unveiling the Timbre Landscape: A Layered Analysis of Tenor Saxophone in RAVE Models. *Proceedings of the 21st Sound and Music Computing Conference (SMC)*, 523–530. DOI: [10.5281/zenodo.14362599](https://doi.org/10.5281/zenodo.14362599)
- Carvalho, N., Sousa, J., Portovedo, H., & Bernardes, G. (2025). Exploring Timbre Latent Spaces: Motion-Enhanced Sampling for Musical Co-Improvisation. *International Journal of Performance Arts and Digital Media*, 21(2), 335–361. DOI: [10.1080/14794713.2025.2536409](https://doi.org/10.1080/14794713.2025.2536409)
- Carvalho, N., Sousa, J., Bernardes, G., & Portovedo, H. (2025). Motiv: A Dataset of Latent Space Representations of Musical Phrase Motions. *Proceedings of Audio Mostly 2025 (AM.ICAD25)*, 82–93. DOI: [10.1145/3771594.3771602](https://doi.org/10.1145/3771594.3771602)

We have also presented this research as a communication at:

- Carvalho, N., Sousa, J., Bernardes, G., & Portovedo, H. (2025). *Echoes of the Unseen: rethinking the performative role of tape through AI-mediated improvisation*. *Comunicação em Digital and Technological Performativities: Book of Abstracts of the 1st International Colloquium*, Aveiro.
- Carvalho, N. & Sousa, J. (2025). *Echoes of the Unseen: improvisation, latent timbres, and Human-AI interaction in real time*. *Book of Abstracts of the XIV Encontro de Investigação em Música (ENIM25)*. DOI: [10.5281/zenodo.17648883](https://doi.org/10.5281/zenodo.17648883)

1.4.3.2 Research Datasets

The following datasets provide the empirical grounding for the latent spaces utilized in the software systems, focusing on the alignment of timbre and musical gesture.

- **SaxTimbre: Tenor Saxophone Extended Techniques** (Carvalho and Sousa, 2024): A comprehensive dataset of saxophone timbres designed for training RAVE models and analyzing timbre landscapes. DOI: [10.7910/DVN/WSFBIJ](https://doi.org/10.7910/DVN/WSFBIJ)

- **Motiv: Musical Phrase Motion Dataset** (Carvalho and Sousa, 2025b): A collection of latent space representations that capture the motion and trajectory of musical phrases, enabling motion-aware sampling strategies. DOI: [10.7910/DVN/RWCG4B](https://doi.org/10.7910/DVN/RWCG4B)

1.4.3.3 Software and Technical Repositories

The software developed for this thesis focuses on real-time interactivity and the navigation of high-dimensional manifolds.

- **BroadcastJSB** (Carvalho, 2026a): A web-based application (Vue3 and Tone.js) that implements a radio receiver metaphor for navigating the latent space of J.S. Bach's chorales. It translates high-dimensional coordinates into audible harmonic structures.

Repository: github.com/NadiaCarvalho/BroadcastJSB

Interactive Tool: nadiacarvalho.github.io/BroadcastJSB/

Related Research: (Carvalho and Bernardes, 2023, 2024b,a).

- **Aethra** (Carvalho, 2026b): An autonomous musical agent built using Pure Data and Python. It utilizes a co-creative loop to manage contextual navigation between a performer and a generative timbre model.

Repository: github.com/NadiaCarvalho/Aethra

Related Research: (Carvalho et al., 2024, 2025b,a).

We also developed two auxiliary web-applications related to Chapters 5 and Chapter 7:

- **Latent Tonal Music Encodings**: An interactive web-platform that allows users to explore the latent behavior and geometric properties of these encodings across a diverse range of symbolic music compositions, providing a tangible bridge between the mathematical representations and their musical outcomes.

Repository: github.com/NadiaCarvalho/Latent-Tonal-Music/tree/website-encodings

Interactive Tool: nadiacarvalho.github.io/BachLatentNavigator/

Related Research: (Carvalho and Bernardes, 2023).

- **Bach Latent Navigator**: An interactive evaluation platform developed to visualize and test the sampling strategies discussed in Chapter 7. It provides a coordinate-based visualization of the latent space, enabling the comparative analysis of linear interpolation versus k -nearest neighbor sampling in preserving tonal logic.

Repository: github.com/NadiaCarvalho/BachLatentNavigator

Interactive Tool: nadiacarvalho.github.io/BachLatentNavigator/

Related Research: (Carvalho and Bernardes, 2024a).

The research is supported by a suite of Google Colab Notebooks that function as interactive, code-first appendices to ensure scientific transparency and reproducibility. These executable documents allow researchers to reconstruct experiments, visualize the high-dimensional latent spaces discussed in the text, and directly interact with pre-trained models without the need for local hardware configuration. By providing these tools, the work bridges the gap between theoretical musical analysis and computational implementation, offering a verifiable methodology for the study of mutable musical works.

- (Carvalho and Bernardes, 2023): [Latent_Tonal_Music.ipynb](#)
- (Carvalho and Bernardes, 2024b): [MCM24_Latent_Spaces.ipynb](#)
- (Carvalho and Bernardes, 2024a): [Bach_Latent_Navigation.ipynb](#)
- (Carvalho et al., 2024): [Unveiling_Sax_Timbre.ipynb](#)
- (Carvalho et al., 2025b): [Sax_Timbre_Navigation.ipynb](#) and [Motions_Experiment.ipynb](#)
- (Carvalho et al., 2025a): [Motiv_Construction.ipynb](#) and [Motiv_Analysis.ipynb](#)

1.5 Structure of the thesis

To address the research questions and objectives previously outlined, this thesis follows a trajectory from the theoretical and symbolic to the sub-symbolic and performative. The subsequent chapters are organized into three main parts, reflecting the progression of the research from its ontological foundations to the development of dynamic, mutable co-creative interfaces.

- **Part I: Framing the Mutable Musical Work** establishes the theoretical grounding for the mutable musical work. Building upon the concepts introduced in this chapter, it explores the shift from the symbolic openness of the 20th century to the high-dimensional latent spaces of current generative models. It situates the work within the context of the digital score as a sentient environment, providing the philosophical justification for re-framing the performer as a navigator.
- **Part II: Latent Spaces of Tonal and Structural Dimensions** focuses on the structural and formal logic of the musical work. It addresses the technical transition from symbolic notation to high-dimensional state spaces. This part explores the mathematical modeling of musical form by aligning variational autoencoders (VAEs) with formal tonal logic, demonstrating how computational openness is structured within a latent space. Finally, it details the development of BroadcastJSB, a practical application that implements a navigation-based interaction model for the real-time exploration of tonal manifolds, presenting an alternative to traditional score-following with the exploration of latent musical landscapes.

- **Part III: Timbre, Gesture, and Motion in Latent Spaces:** Investigates the performative dimensions of contextual navigation. Through the exploration of the RAVE architecture and the development of the Aethra system, this part examines how sub-symbolic timbre landscapes and motion-aware datasets transform the performer into a navigator of machine agency.
- **Part IV: Towards a Co-Creative Practice:** Synthesizes the theoretical and practice-based findings. It evaluates the success of the mutable interface in bridging the gap between the machine's contextually bounded possibilities and expressive control, situating the research contributions, specifically the concept of contextual navigation, within broader musicological and technological discourses on distributed authorship.

Through this narrative, the thesis demonstrates how emerging computational techniques can extend compositional and performative possibilities, offering both practical tools and theoretical frameworks for understanding the mutable musical work.

AI Disclosure Statement

The author acknowledges the use of generative artificial intelligence and AI-assisted tools in the preparation of this thesis to support structural organization and linguistic clarity. The specific applications were as follows:

- **Structural Development and Synthesis:** The thesis's high-level organization was aided by Gemini (Google, 2026). In order to create smooth transitions between chapters, this involved synthesising the author's own previously published papers and research notes. In order to guarantee a logical flow of arguments, the tool was also used to compile and classify preliminary research notes about reviewed literature.
- **Linguistic and Stylistic Refinement:** Grammarly (Grammarly, 2026) was utilized as a technical editing tool to improve readability, correct grammatical errors, and keep a consistent academic tone throughout the manuscript.

All primary research, data collection, and original analysis were conducted independently by the author. Although AI tools helped with gathering information and organizing ideas, everything produced was thoroughly checked, cross-referenced with original sources, and edited by the author. Ultimately, all the conclusions and intellectual contributions in this work are their own responsibility.

Chapter 2

Towards a Mutable Musical Work: A Literature Review

The mutable nature of the musical work invites a reconsideration of how music is conceptualized, created, and experienced in technologically mediated contexts. To frame this investigation, it is necessary to trace the historical and theoretical foundations that inform current understandings of composition, performance, and co-creation within the domains of electroacoustic and computer music.

Rather than a chronological survey, this chapter synthesizes the evolution of musical practice alongside shifting theories of artistic agency. The chapter follows two primary trajectories. First, in Section 2.1, we examine the evolution of compositional practices within electroacoustic and computer music, tracing the shift from material inscriptions (tape and rule-based algorithms) to deep learning paradigms. Second, in Section 2.2, we analyze how these technologies redefine authorship and agency, framing the musical work as a mutable, mediated interface. These perspectives highlight the growing role of computational systems as mediators in the creative dialogue between composers and performers, hoping to establish the conceptual foundation for viewing musical works as mutable interfaces.

2.1 Evolution of Practice: From Inscription to Generative Systems

The evolution of electroacoustic and computer-assisted music over the 20th and 21st centuries showcases an ongoing conversation between music-making and technology (Roads, 2015; Manning, 2013). These developments have brought us from physical manipulation of sound media to operating complex artificial intelligence systems (Magnusson, 2019; Armitage et al., 2024a). Not only have these changes expanded the range of sounds available to composers, but they have also transformed how musical structures are formalized, represented, and executed (Roads, 2004, 2015). This literature review explores this evolution through four main stages: the material and signal inscriptions of the tape era, the rise of symbolic rule-based algorithms, the development of interactive systems, and the current focus on deep learning technology.

2.1.1 Material Inscriptions: From Tape to Studio-as-Instrument

Early electroacoustic practices, emerging in the mid-20th century, laid the foundation for integrating digital technology into compositional practice. The complementary traditions of *Musique Concrète* (Schaeffer, 1998), *Elektronische Musik* (Eimert, 1957; Stockhausen, 1957), and the United States tape music movement (Manning, 2013; Solomos, 2020) articulated divergent philosophies of sound organization. The first, grounded in the phenomenology of recorded sound, relied on reduced listening (*écoute réduite*) to isolate the *objet sonore* (sound object) from its physical cause (Chion, 1983). The second focused on the abstract manipulation of synthesized sine tones, seeking total serial control over every parameter of the acoustic material (Manning, 2013). The third, initially developing without the major institutional support seen in Europe, fractured into two distinct camps: an avant-garde faction focused on radical tape manipulation and indeterminacy, and a more conservative “Tape Music” movement that utilized the medium to extend traditional tonality and pioneered the integration of prepared tape with live instrumental performance (Manning, 2013).

Both European traditions moved beyond the physical constraints of the Western traditional musical staff notation toward direct material engagement with signal inscriptions (e.g., the physical waveforms inscribed on magnetic tape) (Magnusson, 2019). This prompted a transition toward alternative descriptive systems, such as sonograms and algorithmic scripts, relocating the performative act into the fabric of the compositional process. Schaeffer (1966)’s *Solfège de l’objet sonore*, described in his *Traité des objets musicaux* taxonomy, replaces the homogeneous pitch-duration grid with descriptive, morphological parameters like grain, mass, and dynamics. In this context, tape editing became a performance of composition, where the composer acted as a sculptor of sounds. The physical manipulation of the medium (i.e., cutting, reversing, and looping tape) represented a move from abstract symbolization to direct material engagement (Roads, 2015).

Conversely, the German school of *Elektronische Musik* approached the electronic medium through the lens of integral serialism, seeking to extend strict mathematical control to elements of musical creation that had previously been considered volatile or difficult to quantify, most notably timbre and space (Manning, 2013; Solomos, 2020). Led by figures such as Herbert Eimert and Karlheinz Stockhausen at the WDR studio in Cologne, composers utilized sine wave oscillators to build sounds from basic frequencies, allowing them to systematically compose timbre via additive synthesis rather than merely accepting the innate characteristics of acoustic instruments (Manning, 2013; Solomos, 2020). As Eimert asserted, the goal of this electronic serial music was to ensure that “everything to the last element of the note is subjected to serial permutation” (Manning, 2013). This parametric organization naturally extended to the spatial projection of sound, transforming representational space into a fully composable vector (Solomos, 2020). In works such as Stockhausen’s *Gesang der Jünglinge* (1956) and *Kontakte* (1958–1960), the exact localization, rotation, and movement of sound masses across multiple loudspeakers were meticulously serialized and integrated into the formal structure of the composition (Schedel, 2007; Roads, 2015). Ultimately, this paradigm demonstrated that the technological medium could elevate timbre and space from

secondary acoustic byproducts to fully controllable, structural parameters (Roads, 2015).

Concurrently, the tape music movement in the United States initially emerged without the major institutional backing enjoyed by its European counterparts (Manning, 2013), leading to a bifurcation into two distinct aesthetic camps. On one hand, the avant-garde Music for Magnetic Tape project, formed by John Cage in 1951 at the private studio of Louis and Bebe Barron, embraced radical tape manipulation, chance operations, and indeterminacy (Manning, 2013). Compositions such as Cage's *Williams Mix* (1952–1953) were painstakingly assembled from over 500 tiny spliced fragments of recorded sound, illustrating a desire to explore complex sonic montages rather than traditional musical forms (Solomos, 2020). Conversely, the official Tape Music movement, led by Vladimir Ussachevsky and Otto Luening, adopted a more conservative philosophy, treating the tape recorder primarily as a means to extend traditional concepts of tonality and instrumentation (Manning, 2013). This latter school pioneered the integration of prepared magnetic tapes with live instrumental performance, securing commissions for major ensembles with hybrid orchestral works such as *Rhapsodic Variations* (1954) (Manning, 2013). Their successes eventually attracted vital institutional support, culminating in the establishment of the highly influential Columbia-Princeton Electronic Music Center (Manning, 2013; Solomos, 2020), where composers like Milton Babbitt utilized the programmable RCA Mark II Synthesizer to achieve precise, mathematically sequenced control over electronic events (Manning, 2013; Roads, 2015).

As electroacoustic practices developed, the rigid boundary between composition and performance began to blur. In the analogue studio, the physical actions of playing tape recorders, manipulating filters, and executing fader gestures on a mixing desk became central to the creative act (Roads, 2015). Early pioneers like Pierre Schaeffer envisioned this comprehensive setup as an enormous instrument, birthing the concept of the studio-as-instrument (Emmerson, 2007). Consequently, the studio was no longer just a place to capture fixed artifacts, but a highly performative site where the composer acted as a real-time performer of the equipment. This physical, interpretative dimension of tape composition naturally sought a public outlet, prompting composers to literally take the studio out on stage (Emmerson, 2007). By the 1960s, the very tools of the studio were reinvented as live performance instruments, effectively bridging the gap between tape music and live electronics (Schedel, 2007). For example, Pauline Oliveros removed magnetic tape from its traditional recording context and utilized massive tape-delay systems as live performance instruments to build complex, accumulating textures in real-time, as exemplified in works like *I of IV* (Roads, 2015). Expanding upon this hands-on ethos, composers such as David Tudor treated the electronic circuit itself as a performative site. Tudor, particularly in works like *Rainforest*, pioneered the concept of “composing inside electronics,” where custom-built circuits and vibrating physical objects effectively became the score itself (Kuivila, 2004; Collins, 2007b). The work emerged from the idiosyncratic, emergent behavior of the hardware, fully blurring the lines between a fixed artifact and an unpredictable live event, setting the stage for the computational mutability that followed (Sallis et al., 2018).

2.1.2 Formalized Logic: Algorithmic Foundations and Meta-Composition

As electronic computers became available in the 1950s, compositional practice expanded to include algorithmic and generative systems, moving from physical tape to symbolic code. [Hiller and Isaacson \(1979\)](#)'s *Illiad Suite* (1956) pioneered this approach, using rule-based heuristics to generate material through Markov chains, the Monte Carlo method, and formal 16th-century counterpoint logic ([Roads, 1996](#)).

During this era, composition was heavily conceptualized through formal logic and mathematical processes. Algorithms were generally divided into two categories: deterministic algorithms, which executed fixed, rule-based tasks using specific seed data without random selection, and stochastic procedures, which integrated random choices weighted by probability tables ([Roads, 1996](#)). [Xenakis \(1992\)](#) surpassed manual melodic intuition with his Stochastic Music Program (SMP), used to compose his *ST series*. By employing stochastic processes, such as Poisson distributions and Maxwell-Boltzmann kinetic theory, Xenakis utilized the computer to synthesize the internal behavior of sound masses, freeing himself from tedious bookkeeping to act instead as a pilot guiding sonic galaxies ([Manning, 2013](#); [Essl, 2007](#)).

Conversely, [Koenig \(1970\)](#) replaced strict serial permutations with aleatoric principles in his Project 1 software. Rather than modeling physical phenomena like Xenakis, Koenig applied a series of selection principles to a database of musical parameters (instrument, rhythm, harmony, register, and dynamics). By encoding an algorithm that mapped various states between absolute regularity and irregularity, Koenig automated the organization of dense structural elements, firmly shifting the composer's role toward algorithmic design and the curation of generated possibilities ([Essl, 2007](#)).

While [Koenig \(1970\)](#) focused on procedural abstraction, [Cope \(1996\)](#) introduced stylistic recombinant logic with Experiments in Musical Intelligence (EMI). By utilizing augmented transition networks (ATNs) to analyze and deconstruct the stylistic DNA of classical works, Cope shifted the focus toward the algorithmic modeling and simulation of specific musical identities ([Roads, 1996](#)).

This computational transition led to the concept of meta-composition. Instead of writing individual notes, the composer's focus shifted toward defining high-level attributes (e.g., density, speed, and register) that drive the algorithmic procedures responsible for generating musical surface events ([Roads, 2015](#)). A prime example of this is [Truax \(1978\)](#)'s Poisson Distribution (POD) system, which utilized tendency masks to define global frequency-versus-time boundaries within which thousands of micro-level sound events could be stochastically distributed ([Roads, 1996](#)).

A defining limitation of these early rule-based systems was their batch-processing mode of interaction ([Manning, 2013](#)). The composer formulated software away from the machine, fed initial seed data and rules into a mainframe computer using punched cards, and waited hours or even weeks for the job to execute. The composer then manually transcribed the resulting alphanumeric printouts into traditional musical scores ([Manning, 2013](#); [Roads, 1996](#)). Because there was absolutely no real-time interaction or auditory feedback, the algorithm served as a rigid, autonomous

formal engine.

Furthermore, as these generative programs evolved into real-time interactive music systems, they fostered a highly diluted notion of the instrument, the interface, and the score (Magnusson, 2019; Armitage and Magnusson, 2023). In these entangled digital environments, the distinction between composing the work and designing the instrument essentially disappears (Magnusson, 2019). The score morphed from a static timeline of prescriptive events into the structural code of the instrument itself, a dynamic collection of rules, constraints, and interactive mappings (Armitage and Magnusson, 2023). Consequently, these compositions assume a radically open nature. Because the music is generated dynamically in real-time based on these underlying rules, the piece renders differently every single time it is run (Essl, 2007). A prime example of this is Karlheinz Essl's *Lexikon-Sonate* (1992–2010), an infinite interactive composition environment that algorithmically generates piano music on the fly, guaranteeing a unique structural iteration during every performance (Essl, 2007). This paradigm redefined composers as designers of fluid, emergent systems rather than creators of fixed pieces, definitively establishing the computer as a formalizing agent capable of executing complex structures that far exceeded the manual capacity of human composers.

In the context of *musique mixte*, this computational capacity facilitated real-time interactive systems that fundamentally altered the role of the performer (Sallis et al., 2018). A prominent example is Philippe Manoury's *Jupiter* (1987) for solo flute and live electronics, the first piece in his *Sonus ex machina* cycle (Rowe, 2003; Solomos, 2020). Rather than requiring the instrumentalist to rigidly synchronize with a fixed, prerecorded tape, the piece utilizes a computer to continuously analyze the live audio signal captured by the performer's microphone (Sallis et al., 2018). The system then uses this performance data to dynamically control digital signal processing algorithms, generating real-time electronic transformations of the live acoustic input (Rowe, 2003). Consequently, the role of the performer shifts from simply executing a predetermined sequence to engaging in a responsive, mutual partnership with the machine (Sallis et al., 2018). Professional instrumentalists frequently compare the experience of performing these interactive solo pieces to playing traditional chamber music (Féron and Boutard, 2018). Because the electronic segment reacts dynamically to the acoustic input, the performer must continuously listen to the system's output and adjust their physical gestures to influence the machine's subsequent behavior (Féron and Boutard, 2018). This paradigm effectively treats the technological system as an active ensemble partner, requiring the interpreter to adapt to the electronics just as they would to another human musician (Féron and Boutard, 2018).

By the late 20th century, the limitations of early batch-processing were superseded by powerful Computer-Assisted Composition (CAC) tools and programming environments. The rise of these interactive algorithmic systems fundamentally altered the way music and computation interact (Nierhaus, 2009). As computational power extended, platforms like Common Music (Taube, 1991), Common Lisp Music (Schottstaedt, 1994), PatchWork (Laurson and Duthen, 1989), and OpenMusic (Assayag et al., 1999) allowed composers to manipulate and refine their algorithms directly. This change transformed composition into a highly interactive process, encouraging ex-

ploration and negotiation that aligns with modern ideas of co-creation and adaptive systems (McCormack and d’Inverno, 2012). In this paradigm, composers began employing algorithms not just to generate notes, but to explicitly explain their creative choices, providing structural documentation that facilitates replication, analysis, and iterative refinement (Roads, 2015). Consequently, technology begins to operate as a co-creative agent, in which the composer defines the probabilistic constraints, and the system enacts a unique iteration of the work, blurring the line between the pre-determined score and real-time structural evolution.

The evolution of these computational environments prompted composers to explore new architectures of musical form, oscillating between top-down planning, where global macro-structures dictate surface-level micro-events, and bottom-up emergence, where simple algorithmic rules generate complex, unpredictable mesostructures (Roads, 1996). A pivotal milestone in this transition was Hiller and Baker’s MUSICOMP, widely considered the first algorithmic software environment, which provided composers with an abstracted toolkit for probability distributions and tone row manipulations (Roads, 1996). However, as these algorithms expanded to include evolutionary and genetic systems, they encountered severe human-computer interaction limits. Systems that relied on human evaluation to select the best generative outputs suffered from the fitness bottleneck. The overwhelming amount of listening required to evaluate large populations of generated melodies quickly caused user fatigue, effectively detuning the evolutionary effectiveness of the system (Kaliakatsos-Papakostas, 2021).

2.1.3 Interactive and Agential Systems

The rigidity of batch-mode algorithmic composition eventually gave way to interactive paradigms. Instead of viewing the artistic work as a static, finished product, the focus decisively shifted to seeing it as an emergent, continuous process (Shanken, 2002; Di Scipio, 2003). Early hybrid systems, such as Max Mathews’s GROOVE at Bell Labs, embraced this philosophy by allowing composers to modify parameters during playback, shifting the computer’s role from a delayed calculating engine to a live, responsive performance instrument (Mathews and Moore, 1970; Roads, 1996).

Rowe (1993) formalized this shift with his definition of Interactive Music Systems (IMS) as computer programs that participate in live performance by actively changing their algorithmic behavior in response to incoming musical input (Rowe, 2009). To categorize the diverse architectures that emerged, Rowe proposed a foundational taxonomy organized along three operational dimensions (Rowe, 1993). First, he distinguished between score-driven systems (which track a predetermined musical timeline) and performance-driven systems (which react spontaneously to unscripted input). Second, he classified their response methods as transformative (altering live or existing musical material), generative (using algorithmic rules to produce new material from elementary seeds), or sequenced (triggering pre-recorded fragments). Finally, he delineated the instrument paradigm, where the system acts as an extended performance controller, from the player paradigm, where the computer operates as an independent musical partner (Rowe, 1993).

A quintessential realization of Rowe’s score-driven category is found in the extensive machine listening research conducted at the Institute for Research and Coordination in Acoustics/Music

(IRCAM). Tools such as Antescofo (Cont, 2008) operate as highly advanced score followers, continuously analyzing the live audio stream of human performers to dynamically synchronize complex electronic actions with a predetermined musical score (Nika, 2016).

Conversely, the performance-driven approach facilitated the rise of Interactive Machine Learning (IML) and generative performance systems. Tools like the Wekinator (Fiebrink and Cook, 2010), for example, allow artists to bypass explicit programming and instead create complex, transformative, and generative mappings between physical gestures and sound synthesis by providing the computer with small data sets of input-output examples, prioritizing a human-in-the-loop exploratory practice (Fiebrink and Sonami, 2020; Vigliensoni et al., 2022). Extending this exploratory control beyond physical gestures, frameworks like the “composition of processes” (Bouche et al., 2016) apply dynamic mutability directly to the score, treating it as a continuously evolving program rather than static data.

Building on this ecological approach to interaction are agential scores, which use artificial life to create fluid, emergent, and self-organizing musical notations (Armitage and Magnusson, 2023). Utilizing software libraries like Tölvera alongside IML frameworks like Anguilla, composers establish environments driven by what Armitage and Magnusson term “intra-active perturbations” (Armitage et al., 2024a). Drawing on Karen Barad’s philosophy of agential realism, this framework suggests that human and non-human agencies do not pre-exist as separate entities. Rather, they are mutually constituted through their entanglement (Armitage and Magnusson, 2023; Armitage et al., 2024a). This marks a deliberate departure from the traditional learner-teacher paradigm of machine learning, instead fostering a continuous, non-hierarchical dialogue of mutual provocations between the performer, the instrument, and the score.

2.1.4 Deep Learning in Compositional Practice: The Democratization of Music

The most recent shift in compositional practice marks a transition from symbolic logic and expert systems to deep learning (DL) and machine listening paradigms (Magnusson, 2019). Unlike rule-based algorithms, where programmers explicitly define compositional syntax, DL systems (e.g., Google’s Magenta or DeepMind’s WaveNet) employ artificial neural networks that learn autonomously from massive datasets (Briot and Pachet, 2020). This paradigm fundamentally alters the composer’s relationship with the machine. Deep symbolic models, utilizing Recurrent Neural Networks (RNNs), variational autoencoders (VAEs), and Transformers, process music as discrete tokens, such as MIDI events, to predict subsequent structures (Ji et al., 2023; Briot and Pachet, 2020). Conversely, systems like WaveNet (Oord et al., 2016) analyze and synthesize audio at the micro-level of individual sample points. More recent models, such as NSynth (Ardaillon, 2017) and RAVE (Caillon and Esling, 2021), further push this boundary. By autonomously discovering acoustic features rather than relying on hand-crafted parameters, these systems allow for the interpolation of entirely novel timbres (Wyse, 2017). The advent of Large Language Models (LLMs) and massive generative architectures (i.e., OpenAI’s Jukebox (Dhariwal et al., 2020) and Google’s MusicLM (Agostinelli et al., 2023)) has facilitated a democratization of musical creation (Park, 2025). These systems can learn complex stylistic markers, timbral nuances, and structural

patterns from vast datasets, allowing users to generate high-fidelity audio from simple textual prompts. This embodies a paradigm shift in user interaction that [Campo \(2014\)](#) describes as “lose control, gain influence,” in which low-level, fine-grained actions are delegated to the machine’s automation.

Despite this mainstream accessibility, these architectures face a fundamental hurdle in computational creativity. Because DL relies on the statistical distribution of training data, it excels at interpolation (blending and imitating existing styles). Still, it fundamentally struggles with extrapolation (the creation of true outlier patterns and novel conceptual leaps) ([Ji et al., 2023](#)). Furthermore, deep learning introduces a profound opacity into the creative process. Because a system’s knowledge is distributed across multidimensional neural connections, scrutinizing the source code alone can no longer explain the resulting output ([Magnusson, 2019](#)). This creates a black box effect that distances the composer from the underlying structural logic. In electroacoustic music, this presents a specific tension: while these models offer powerful synthesis, they often lack the precise, sculptural control inherited from the tape era.

To navigate this opacity, researchers must move toward a practice of interactive, human-centered co-creation and navigable generative environments. For instance, [Fiebrink and Cook \(2010\)](#) advocates for IML, using tools like Wekinator to allow artists to iteratively train expressive models with a “small data mindset” ([Vigliensoni et al., 2022](#)). These systems allow bypassing explicit programming and keeping the human firmly in the generative loop ([Fiebrink and Caramiaux, 2018](#)). In the sub-symbolic audio domain, [Tatar et al. \(2021\)](#) does this by developing Latent Timbre Synthesis (LTS), which utilizes VAEs to map audio corpora into continuous latent spaces. Rather than generating fixed offline outputs, this system allows artists to continuously search, sculpt, and interpolate entirely novel timbres in real time.

To bridge the gap between offline generation and live performance, [Vear \(2022a\)](#) proposes Embodied AI, designing intelligent systems whose operational behavior interacts dynamically as a co-creative partner inside the spontaneous “flow of musicking.” Finally, [Armitage et al. \(2024b\)](#) addresses the interaction gap through agential scores. By commingling artificial life libraries (like Tölvera) with IML, they foster a continuous, non-hierarchical dialogue of mutual provocations between the human performer and the generative system ([Armitage et al., 2024a](#)).

Building upon these interactive paradigms, this thesis moves in the opposite direction of current massive model trends. Rather than utilizing DL for automated generation, this research seeks to open these models, leveraging their internal representations (i.e., the latent spaces) as sites for precise, performative intervention. We view DL not as a black-box end-product generator, but as a rich repository of sonic features that can be navigated and sculpted in real time.

2.2 Agency, Authorship, and the Theory of Mediation

The integration of technology into musical practice alters the traditional roles of the composer, performer, and audience by introducing the computational system as an active participant ([Lattour, 2005](#); [McCormack and d’Inverno, 2012](#)). This section examines how these developments

redistribute authorship and redefine musical agency among these interacting entities. First, the composer's role shifts from creating fixed musical artifacts to designing algorithmic processes, thereby delegating real-time generative decisions to the computational system and the player (Rowe, 1999; Féron and Boutard, 2018). Second, the traditional performer evolves into a “post-performer,” moving beyond executing predetermined notation to function as a co-creator who engages in bidirectional interaction with the machine (Leman, 2007; Magnusson, 2019). Third, the technological system operates as an autonomous or semi-autonomous agent capable of analyzing input and generating musical responses, rather than functioning merely as a passive tool (Rowe, 1993; Collins, 2007a). Finally, the listener transitions from a passive consumer to an active participant, or “visitor-musicant,” who influences and co-authors the musical output within interactive installations (Tiffon, 2012). Ultimately, these transformations shift musical creation from hierarchical, human-centric models toward distributed, ecosystemic networks of agency (Di Scipio, 2003; Waters, 2007).

2.2.1 Authorship and Distributed Agency

Contemporary authorship is becoming fragmented across training data, system design, and real-time interaction. This collective authorship mirrors mid-20th-century experimental art, where composers like Cage, Brown, and Cardew reduced central control to allow for performer influence through indeterminacy (Nyman, 1999).

However, with computational mediation, this redistribution extends beyond human performers, positioning algorithmic systems as active participants rather than passive tools. According to the theory of mediation (or mediology), technical mediums are not neutral conduits, actively transforming how music is conceived, conserved, and transmitted (Tiffon, 2018). Digital instruments and algorithms function as “epistemic tools,” artefacts that engage with musical theory, design biases, and historical knowledge, directly shaping the composer's cognitive process (Magnusson, 2009).

Consequently, authorship emerges from a complex network of interactions among people, computers, and context (McCormack, 2012). To understand this entanglement, researchers increasingly draw upon Actor-Network Theory (ANT) and Barad (2003)'s concept of “intra-action.” Unlike traditional interaction, which assumes independent entities that simply react to one another, intra-action suggests that human and non-human agencies do not pre-exist as separate elements. Rather, they are mutually constituted through their performative entanglement (Armitage and Magnusson, 2023). In this light, the technological system acts as a true co-performer (Granzow, 2018), possessing its own material and computational agency.

As technological mediation fragments traditional authorship into a collective network, it reflects Braidotti (2013)'s posthumanist framework, which argues that our technologically mediated societies displace the “traditional humanistic unity of the subject” in favor of flexible, hybridized identities. This hybridized identity is practically realized in the role of the Computer Music Designer (or *Réalisateur en Informatique Musicale*, RIM) (Zattra, 2018a). The RIM acts as a crucial co-author who translates abstract musical concepts into operative digital facts, meaning that the

technological and aesthetic identity of the work is negotiated through hidden, oral collaborations long before it reaches the stage (Zattra, 2018a). By delegating generative responsibility to these algorithmic and technological processes, the composer shifts to a “meta-level of design” (Rowe, 1999). Each mediating layer, from the algorithm’s code to the performer’s physical movement, adds a layer of authorship to the final result.

Ultimately, these practices challenge the traditional Western Art work-concept, the idea of a fixed, authoritative score. Instead, music becomes an invention: a dynamic locus of behaviors, rules, and relationships that render uniquely in every performance (Magnusson, 2019). The composer acts as a meta-designer, creating frameworks where machine, technician, and performer collaborate, aligning perfectly with post-structuralist views on the “author-function” (Foucault, 1977), where the author is no longer a fixed origin point, but a variable distributed across the entire techno-social discourse.

2.2.2 From Artifact to Process: The Ecology of Digital Preservation

While the conceptual shift toward fluid, interactive, and autonomous systems radically expands musical possibilities, it simultaneously introduces a severe material vulnerability. If the identity of the mutable work resides not in a static paper score but within a complex, digitally mediated performance ecosystem, its survival is entirely dependent on the stability of that technology. This introduces a profound ontological challenge: if the software breaks, the operating system migrates, or the hardware becomes obsolete, does the work die?

Unlike traditional acoustic instruments, which have benefited from centuries of standardized lutherie and established notational practices, digital performance environments are highly unstable (Magnusson, 2019). As Rowe (2009) observes, compositions relying on specific commercial hardware or bespoke software face rapid technical obsolescence. In exploring the long-term preservation of these works, Pennycook (2008) highlights a severe vulnerability: when complex, custom-built technological systems fail or become obsolete, the interactive dimension is often lost entirely, leaving little more than the written scores and fixed audio recordings as the only surviving long-term artifacts. Because new instruments and interactive systems resemble fluid processes rather than solid states, a piece composed for a specific digital configuration today may be entirely unrealizable in a few years unless meticulous actions are taken to preserve the exact state of the hardware, software, and protocols (Magnusson, 2019). Consequently, without comprehensive preservation strategies or the direct intervention of the composer to “turn the knobs,” many mixed and interactive works are performed only a handful of times before fading from the active repertoire or becoming fundamentally unplayable (Pennycook, 2008; Féron and Boutard, 2018).

This technological fragility deeply destabilizes the traditional Western work-concept, a paradigm historically predicated on the eternal reproducibility and untouchable permanence of the written score (Born, 2005; Magnusson, 2019). When the underlying code, the interface, and the algorithmic logic collectively constitute the composition, functioning as what Magnusson (2019) refers to as a fluid musical “invention” rather than a static artifact, the obsolescence of the digital medium effectively signifies the death of the piece (Pennycook, 2008; Magnusson, 2019). To combat this

vulnerability, the burden of preservation has shifted from the traditional publication of prescriptive scores to the active, ongoing documentation of technological procedures. Ensuring the survival of these interactive works requires composers and computer music designers to meticulously record software patches, hardware configurations, and operative methodologies, creating a comprehensive technological record that allows the performance ecosystem to be reconstructed as digital platforms continue to migrate (Pestova, 2018; Zattra, 2018b).

This ontological fragility further cements the role of the RIM as a collaborator in the creative process. In the context of preservation, the RIM's role extends to that of a digital archivist. Because standard notation cannot capture the behavioral nuances of live electronics or machine learning models, preservation relies heavily on documenting the technological setups, software patches, and operative methodologies to ensure that the work can be migrated and reconstructed as technologies inevitably age (Zattra, 2018a). Therefore, maintaining the "living score" (Vear, 2019) requires continuous human intervention, proving that the very technology that enables autonomous musical agency is simultaneously the source of its greatest impermanence.

2.2.3 Technological Agency and the Mediated Work

Unlike early deterministic tools, modern ML systems can act in unpredictable ways, proposing musical ideas that the human creator may not have explicitly programmed (McCormack, 2012). Theories like Actor-Network Theory (Latour, 2005) and postphenomenology (Ihde, 1990) suggest that agency does not pre-exist as an isolated property, but emerges dynamically from these entangled relationships. Drawing on Karen Barad's concept of intra-action, researchers increasingly view agency not as something a machine possesses, but as a performative enactment distributed across the human, the machine, and the environment (Saunders and Gemeinboeck, 2022; Armitage and Magnusson, 2023). We see these technological systems as actants or skillful participants (Saunders and Gemeinboeck, 2022) that shape musical expression through three main modes of mediation (Ihde, 1990; Magnusson, 2019):

1. **Embodiment Relations:** Technology becomes transparent to the performer, functioning as a quasi-extension of their body. This occurs when the feedback loop is fast and intuitive enough that the musician forgets the interface and focuses purely on the sound, as seen in highly responsive digital instruments (Ihde, 1990; Nijs et al., 2013). In the context of AI, this extends to embodied AI, where the machine attempts to seamlessly participate in the shared, situated flow of live "musicking" (Small, 1998).
2. **Hermeneutic Relations:** Technology creates a connection to a digital landscape, such as a complex latent space or a dynamic visual score. The performer observes and interprets the machine's state, treating the interface as a fluid text to be decoded and navigated (Magnusson, 2009; Verbeek, 2021). For example, agential scores driven by artificial life algorithms exhibit emergent behaviors that prompt performers to constantly identify with, mirror, and attune to the notation in real-time (Armitage and Magnusson, 2023).

3. **Alterity Relations:** Uniquely amplified by AI, the technology is experienced as an independent entity or playful other rather than a mere tool (Vear, 2022a; Ihde, 1990). The system acts as a conversational partner or co-performer, engaging the human in a continuous cycle of mutual provocations and responses without a strict hierarchy (Granzow, 2018).

The balance between embodiment, hermeneutics, and alterity provides a framework for Explainable AI (XAI) in creative systems (Abdul et al., 2018). While the underlying neural weights of deep learning models remain structurally opaque (as established in Section 2.1.4), XAI addresses this limitation at the interface level. By designing visual and haptic representations that are intuitive (embodied), intelligible (hermeneutic), and interactively robust (alterity), the computational system is transformed from an abstract black box into a legible, symbiotic partner (Fiebrink, 2011; Poltronieri, 2022). The performer and the machine enter a cybernetic feedback loop where both cognitive systems continuously tune one another (DiBlasi, 2022). The performer is thus empowered to negotiate with the system's agency, resulting in a truly co-authored musical experience.

2.2.4 Interfaces as Sites of Co-Creation

The integration of computation into musical practice necessitates a fundamental reevaluation of the interface. As Magnusson (2019) observes, an acoustic instrument is an interface bound by physical laws, whereas a digital instrument has an interface that requires the arbitrary and deliberate mapping of human gestures to sound synthesis algorithms. Consequently, the digital interface functions as an epistemic tool, a symbolic and theoretical construction where the designer's compositional intent and embedded music theory dictate the parameters of interaction (Magnusson, 2009). It acts as a boundary object (Armitage and Magnusson, 2023) where performative gesture and algorithmic behavior converge (Fiebrink, 2011). Within this framework, the interaction shifts from purely physical dexterity to an epistemically mediated embodiment, where the performer must dynamically navigate the conceptual and symbolic logic of the computational environment (Magnusson, 2009; Vear, 2019). Understanding this interaction requires analyzing the system's ergonomics, which translates into the comprehensive framework of the system's material affordances, encoded constraints, and the resulting potential for musical exploration, analogous to the concept of gameplay in interactive media (Magnusson, 2019).

2.2.4.1 From Control Surfaces to Mediating Environments

Interaction has evolved from simple one-to-one mappings (e.g., a single fader controlling volume) to digital mediating environments built in languages like Max/MSP, Pure Data, and SuperCollider. These object-oriented platforms facilitate many-to-many mappings, where a single physical gesture can influence dozens of synthesis parameters simultaneously (Hunt et al., 2000). To understand why this gestural complexity is necessary, we must look to Marc Leman's theory of Embodied Music Cognition (Leman, 2007). Leman argues that the human body acts as the natural

mediator between the musical mind and physical sound energy. We process music through corporeal articulations and a continuous action-reaction cycle, engaging in a process of mirroring where moving sonic forms are translated into a subjective, action-oriented ontology (Leman, 2007).

In acoustic instruments, this sensorimotor coupling is innate, driven by direct mechanical causality and haptic feedback. However, because digital instruments physically decouple the gestural control interface from the sound production mechanism, they risk severing this embodied connection (Leman, 2007). By moving away from commercial interfaces designed to be perfectly effortless, designers deliberately introduce material resistance and non-linear unpredictability (Collins, 2007a; Magnusson, 2009). This designed friction influences the performer to adapt and physically negotiate with the system in real-time, effectively restoring the corporeal resonance and motor-mirroring necessary for true musical flow.

In this context, the score shifts from a static set of prescriptive instructions into a dynamic digital environment that performers actively inhabit (Drummond, 2009). A pioneering example of this is Lewis (2000)'s *Voyager*, an interactive system that positioned the machine as a virtual improviser. Operating under the player paradigm, *Voyager* analyzes incoming performance data to condition its own generative responses without strict subservience, establishing a non-hierarchical dialogue where neither the human nor the machine is entirely in charge (Rowe, 1993; Lewis, 2000).

Modern interfaces leverage machine learning and artificial life to prioritize this embodied dialogue over strict commands. For Vear (2019), this manifests as the living score, requiring an embodied state of flow where the performer is constantly discovering the work as they play it (Vear, 2022a). This interactivity is further radicalized by agential scores, such as those built with the Tölvera library, which imbue points and lines with fluid material agency and autonomous, goal-seeking behaviors, prompting performers to attune to the notation as a living entity (Armitage and Magnusson, 2023). Similarly, tools like Wekinator use Interactive Machine Learning (IML) to allow musicians to bypass explicit programming, instead training their own expressive, embodied mappings through rapid input-output examples (Fiebrink and Sonami, 2020). Examples also include corpus-based concatenative synthesis (CBCS) in CataRT (Schwarz et al., 2006), which allows for the fluid, gestural navigation of massive, multi-dimensional sound libraries.

2.2.4.2 Epistemic and Performative Roles

These interfaces serve two leading roles in the co-creative process. First, they act as epistemic tools for knowledge-building (Magnusson, 2009). Rather than acting as neutral conduits, digital interfaces encapsulate specific musical theories, historical knowledge, and design biases. Through visual, auditory, and haptic feedback, the interface reveals the system's internal organization and its ergodynamics (the comprehensive framework of its affordances, constraints, and expressive potential) (Magnusson, 2019). As the performer navigates the interface, they actively internalize the theoretical structure of the work, decoding its probabilistic tendencies and sonic boundaries to construct a functional cognitive model of the machine's underlying logic.

On the other hand, interfaces play a performative role, in the sense that they facilitate an embodied state of creative flow and live musicking (Vear, 2022a; Small, 1998), wherein the boundary between human intentionality and systemic output becomes integrated. Within high-dimensional generative spaces, the interface functions as a site for a “dance of agency” (Pickering, 1995). As (Zappi and Tatar, 2025) observe, neural audio instruments amplify this dynamic by acting as embodied cognitive extensions that fundamentally reshape the performer’s phenomenological and epistemological relationship with the machine. In this context, the system acts as a skillful participant rather than a passive instrument (Saunders and Gemeinboeck, 2022). It autonomously generates contextually bounded sonic proposals, while the human actively selects, refines, or reacts to them, thereby establishing a continuous, non-hierarchical loop of intra-action (Armitage and Magnusson, 2023).

Ultimately, this research conceptualizes the musical interface not merely as a control surface, but as a meta-interface: an intelligent, ecosystemic environment (Di Scipio, 2003; Waters, 2007) that coordinates the dynamic negotiation of agency across all human and non-human participants.

2.2.5 Autonomous Agents and the Topological Turn

The evolution of the digital score into a sentient environment (Vear, 2019) finds its technical realization in autonomous creative agents. This represents a profound shift toward co-creativity, wherein the machine transcends the role of a passive tool to become a collaborative partner with its own stylistic preferences (Assayag, 2021). In this context, true autonomy is defined not merely by generative capacity, but by the system’s ability to evaluate its environment and generate its own internal goals (d’Inverno and Luck, 2004; Collins, 2007a). This is effectively formalized in the criteria for the framework Live algorithms (Blackwell et al., 2012), which state that a machine acting as a true improvisational partner must exhibit four core attributes: autonomy, novelty, participation, and leadership (Blackwell et al., 2012). The inclusion of leadership is a powerful theoretical mechanism. It asserts that the machine does not merely react to human input but actively steers the musical direction, forcing the human to adapt to its suggestions. As seen in frameworks like Arne Eigenfeldt’s Musebots, these autonomous systems interact with human performers as independent musical personalities, engaging in a non-hierarchical social negotiation (Eigenfeldt, 2022).

The navigation of these high-dimensional generative spaces has increasingly relied on topological models. The SomSwarm project (Mogenssen, 2020), for example, improved topological navigation by combining self-organizing maps (SOM) with swarm algorithms to build a dynamic spatial memory of sound features. Similarly, the REACH project’s Somax (Borg, 2019; Borg et al., 2022; Fiorini and Malt, 2023) architecture achieves a floating musical coordination with human improvisers by translating live musical streams into activations across specific zones of a pre-annotated corpus. Utilizing layered architectures, including Listening Modules to analyze live input and Matching Modules to navigate the corpus, the machine continuously adjusts to its surroundings while maintaining a consistent stylistic output.

However, while these advanced systems excel at real-time reactivity, they often face a profound control gap (Assayag et al., 2022; Teixeira et al., 2024). Because interaction is primarily

based on low-level acoustic features or step-by-step statistical matching, the connection between human input and machine logic remains abstract rather than structurally directed. This exposes a failure in what [Roads \(2015\)](#) terms “multiscale organization,” the ability to plan structures across multiple timescales simultaneously, from micro-level grains to macro-level forms. Consequently, these systems frequently lack the compositional foresight or anticipatory behavior required to intentionally navigate idiomatic musical macro-structures, such as harmonic progressions ([Nika, 2016](#)).

This thesis seeks to bridge this gap by explicitly connecting the sub-symbolic domain of audio corpora with the symbolic domain of musical structure ([Rowe, 2009](#)). To achieve this multi-scale coherence, we draw upon [Nika \(2016\)](#)’s “scenario/memory model.” In this framework, the machine’s improvisation is guided by a formal temporal structure (the scenario, such as a chord progression) which acts as a macro-level plan, while drawing upon an indexed audio corpus (the memory) for micro-level generative freedom. By linking system memory to formal tonal logic, music-related taxonomies, and dynamic temporal scenarios, the system is imbued with anticipatory behavior, allowing it to prepare for resolutions and cadence points just as a human musician would ([Nika, 2016](#)). By intertwining [Magnusson \(2009\)](#)’s epistemic tools with [Vear \(2019\)](#)’s digital scores, this research aims to develop a model where performers negotiate with the machine’s autonomous capabilities as intuitively and structurally as they would with a traditional acoustic instrument.

2.3 Summary

By tracing the progression from fixed physical recording media to rule-based algorithmic systems and deep learning, this chapter establishes the technical foundation for adaptable, computationally mediated music. This trajectory demonstrates a shift away from the traditional concept of the musical work as a static artifact toward a model of music as a dynamic, responsive process ([Magnusson, 2019](#); [Vear, 2022a](#)). Instead of a finalized product, the contemporary electroacoustic piece is structured as an interactive network of interdependent human and technological components that continuously adapt to environmental inputs. Within this framework, authorship is distributed across the system, and creative output relies on real-time, bidirectional data exchanges between the human performer and autonomous machines operating without a predefined control hierarchy ([Saunders and Gemeinboeck, 2022](#); [Armitage and Magnusson, 2023](#)).

However, this reliance on adaptable digital systems introduces significant technical vulnerabilities. Because the music depends entirely on specific hardware and software configurations, it is highly susceptible to rapid obsolescence. Therefore, preservation efforts must shift from archiving traditional paper scores to documenting technological setups, operating methodologies, and functional code ([Rowe, 2009](#); [Zattra, 2018a](#)). Instead of creating fully autonomous programs that replace human involvement or quickly become outdated, this framework supports designing understandable computer interfaces that act as collaborative, cognitive instruments, or epistemic tools, to enhance human creativity ([Magnusson, 2009](#); [Poltronieri, 2022](#)).

This theoretical foundation necessitates a shift toward practical frameworks that prioritize both expressive agency and ontological resilience. Consequently, the subsequent chapter of this thesis introduces models that utilize latent space navigation and network-based representations to bridge the identified control gap (Nika, 2016). By explicitly synthesizing sub-symbolic audio depth with symbolic formal logic (Rowe, 2009), the research provides the composer with the structural foresight required to navigate high-dimensional generative environments. This approach ensures that the resulting musical work remains a navigable, sustainable entity, one that preserves the spontaneous, embodied flow of live musicking while grounding the machine's autonomous agency within a legible human dialogue (Small, 1998; Vear, 2022a).

Chapter 3

From Descriptor to Latent Spaces: A Topographical Review of Musical Models

With the shift from prescriptive composition to co-creative exploration, there is a critical need for a robust framework capable of representing musical information, whether pitch and harmonic qualities, rhythm, or timbre. While autonomous agents provide the agency for interaction, it is the underlying descriptor spaces and musical models that supply the topography upon which this interaction occurs. This chapter reviews the mathematical, psychoacoustic, and computational models that organize musical data into these navigable structures.

To outline the transition from discrete data selection to the continuous topological morphing that defines the mutable musical work, this chapter is divided into four main sections. Section 3.1 examines the foundational topologies of tonal pitch, rhythm, and timbre. By connecting complex acoustic data with human perception, these multidimensional spaces transform musical entities into organized, navigable maps. Moving on, Section 3.2 explores deep generative models that uncover emergent topologies directly from musical data, moving beyond traditional feature extraction. Section 3.3 investigates how corpus-based concatenative synthesis (CBCS) encapsulates these descriptor spaces (both traditional and latent) to enable fluid, gestural navigation of sound corpora. Finally, Section 3.4 discusses methodologies for converting these complex, high-dimensional spaces into explainable formats, ensuring the system remains interpretable and collaborative rather than an opaque black box.

3.1 Descriptor Spaces: The Topography of the Navigable Manifold

Paramount to the co-creative paradigm is the capacity to represent musical information in ways that are both computationally efficient and musically meaningful (Leman, 2007). To computationally model music, we must connect complex data with human sonic experience, ensuring the resulting system reflects genuine human creativity (Ferguson and Brewster, 2018). By utilizing

general descriptors for elements like tone, rhythm, and timbre, a musical piece is transformed into an organized map that unifies structure and emotion into a coherent system (Clarke, 1987b).

This thesis adopts the concept of descriptor spaces as multidimensional representations that capture the nuances of sound, structure, and gesture. These spaces allow us to compare, transform, and recombine musical entities across various representational levels (Schwarz et al., 2006). Crucially, descriptors act as the bridge between a system’s listening and matching modules, enabling the machine to recognize and respond to a performer’s gestural intent within a shared coordinate framework (Borg, 2019). This process acts as a crucial link between perception, analysis, and generation, enabling data-driven creative exploration (Schwarz, 2007) as the distance between points represents how similar they are in terms of perception or style, rather than just a mathematical concept (Berenzweig et al., 2003).

Furthermore, descriptor spaces help create what Vear (2019) calls the “musicianship of the environment.” When a digital score is represented as a descriptor-based manifold, performing the music becomes an act of moving through this space. The material resistance of the piece is defined by the density and distribution of its descriptors. Some descriptors are very useful, while others have limited options or mixed results. Descriptors are not just tools for analysis, helping shape a flexible interface. This document will focus on three main areas of descriptor spaces: tonal pitch, rhythm, and timbre. Ultimately, by organizing layouts that reflect human understanding of music and its diverse styles, these spaces transform the interface into a site for intelligent musicianship, where both performer and machine collaborate dynamically (Drummond, 2009).

3.1.1 Tonal Pitch Spaces: The Geometry of Harmony

In the symbolic domain of music, descriptor spaces frequently manifest as tonal pitch spaces, which formalize the theoretical and cognitive relationships between pitches, chords, and keys (Lerdahl, 1988). Rather than viewing tonality as a linear sequence of events, these models treat harmonic progression as movement through a multidimensional geometry, where the spatial distance between points reflects their stylistic and perceptual similarity (Tymoczko, 2011). By representing music topologically, we can treat the act of composing or performing as navigating a continuous map of sound, enabling creators to explore new ideas while staying anchored to a coherent tonal center.

The literature offers several distinct topologies of tonality, summarized in Table 3.1, which vary based on their mathematical or perceptual focus and serve as the navigational foundation for co-creative systems. The next sections will detail each topology.

3.1.1.1 Lattice-Based Pitch Spaces: The Tonnetz and its Computational Extensions

The Tonnetz, also called the tone network, is a well-known model in music theory for describing musical tones. Euler (1739) introduced it in 1739, and Riemann (1880) made it popular in 1880. This model organizes pitches into a hexagonal grid based on perfect fifths and major/minor thirds,

Table 3.1: Typology of Tonal and Pitch Space Models in Music Theory and Computation.

Model Type & Key Authors	Core Representation / Geometry
Interval Lattice <i>Euler (1739); Riemann (1880)</i>	Tonnetz: Hexagonal grid representing perfect fifths and thirds; emphasizes triadic transformation.
Functional Hierarchy <i>Lerdahl (2001)</i>	Algebraic/Metric: The Tonal Pitch Space calculates distances based on the stability of intervals within a key.
Spiral Array <i>Chew (2002)</i>	Helical: Maps pitches, chords, and keys onto a 3D spiral to represent hierarchical pitch relations.
Geometric Orbifolds <i>Tymoczko (2006)</i>	Möbius Strips/Toroidal: Mathematical manifolds where voice-leading is represented as paths in non-Euclidean space.
Fourier Phase Space <i>Quinn (2006); Amiot (2016)</i>	DFT Coefficients: Complex-plane representations of clumpiness or evenness in pitch collections.
Tonal Interval Space <i>Bernardes et al. (2016a)</i>	12D Vectors: High-dimensional interval vectors (TIVs) that represent perceptual distances and tonal tension.

making it exceptionally effective for visualizing parsimonious voice-leading (i.e., chordal transitions with minimal movement between notes) (Tymoczko, 2008), as illustrated in Figure 3.1. Modern Neo-Riemannian theory, using the Tonnetz, explains how late-Romantic composers moved between distant musical keys without using traditional harmony rules. (Cohn, 1997, 1998).

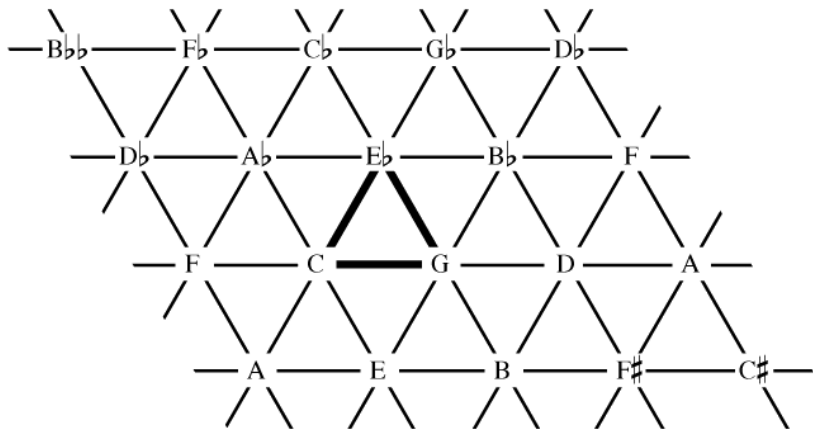


Figure 3.1: Prototypical Tonnetz representation showing triadic configurations. The highlighted triangle represents a C-minor chord triad, demonstrating how the vertical and horizontal axes organize pitch classes into functional harmonic aggregates.

Beyond historical chromaticism analysis, the Tonnetz also provides a strong mathematical basis for modern digital organology and generative systems. When extended into a three-dimensional (3D) toroidal shape (i.e., when pitch classes wrap around both the octave and the cycle of fifths (Tymoczko, 2012)), it allows generative algorithms to explore harmonic space endlessly without encountering mathematical boundaries (Nguyen and Tsabary, 2022), useful for algorithmic path-finding.

In co-creative interfaces, the Tonnetz evolves from a basic diagram into an effective navigation tool. Interfaces like Isochords (Bergstrom et al., 2007) or HexaChord (Bigo and Andreatta, 2016) show how this hexagonal geometry can display real-time performance data, giving the performer a clear view of their harmonic position. By clearly showing these relationships, the interface helps users understand complex harmonies and perform triadic transformations. These transformations include the parallel (*P*), leading-tone (*L*), and relative (*R*) operations, which shift a triad to its parallel, leading-tone, or relative form by moving a single pitch. The distance moved on the interface shows how much the harmony changes. Performers can use this method to understand how different tonal areas relate to each other, turning harmonic sequences into a clear visual experience.

Recently, researchers have started using the Tonnetz as encodings in DL models. Chuan and Herremans (2018) showed that the Tonnetz can be changed into a two-dimensional image that is useful for training Convolutional Neural Networks (CNNs). Aminian et al. (2020) used the Tonnetz geometry as encoding spaces for Long Short-Term Memory (LSTM) networks. Their research shows that this setup helps networks learn long-term musical patterns and connections. By using the Tonnetz as a guide, these models make the music more stable and improve its overall structure. This suggests that the navigational logic of the Tonnetz serves as a critical bridge, aligning the understanding of machine-learning systems with the intuitive practices of the human musicians who work with them (Chuan and Herremans, 2018).

3.1.1.2 Functional Hierarchy: Tonal Pitch Space (TPS)

Moving beyond simple interval distance, Lerdahl (2001)'s Tonal Pitch Space (TPS) quantifies the psychological dimensions of musical tension and stability across a five-level hierarchy. TPS uses a simple math approach to measure the distance between two musical elements, based on their place in a five-level system: the chromatic level, the diatonic level, the triadic level, the fifth level, and the tonic level. In this space, a chord with the same root note but in a different key is further away from the home tonic than a chord in the same key, even if they produce the same pitches. This difference is essential for perceiving structural voice-leading movements in a musical work (Lerdahl, 1988).

TPS enables measuring tonal tension and relaxation by clearly defining the relationships between musical notes. This system calculates the energy of harmonic progressions by looking at the distances between chords, which shows how different they are from one another (Krumhansl, 1990). To calculate this distance, the number of steps moved along the cycle of fifths is added to the number of different pitches that need to be changed to reach the target chord. This approach helps create co-creative agents that work in a goal-oriented musical context (de Haas et al., 2011).

While a simple model may only look for sound similarities, an agent using TPS can prioritize meaningful resolutions. For example, the system can tell the difference between a stable return to the main note (tonic) and a tense move toward a different note. This ability helps the machine suggest musical ideas that match the emotional flow of a piece (Williams, 2007).

Furthermore, TPS introduces the concept of regional distance, which maps the relationship between keys on a multidimensional grid, allowing a co-creative interface to offer the performer a structured representation of tonal hierarchies, where specific paths of modulation represent greater quantifiable distances and structural divergence than others (Randall and Khan, 2010). By embedding these cognitive constraints into a representational space, the system is able to better distinguish the importance of specific parts of a musical sequence (Lerdahl and Krumhansl, 2007).

3.1.1.3 The Spiral Array: Hierarchical Proximity

While the Tonnetz is effective at representing local triadic relationships, and TPS offers the algebraic logic for hierarchical distance, Chew (2002)’s Spiral Array provides a comprehensive framework for global harmonic tracking, translating these relationships into a continuous 3D geometric manifold. This model represents musical entities as points along nested 3D helices. In this geometry, the spatial proximity between a pitch and a key is determined by their tonal function rather than simple frequency. For instance, C and G chords are positioned close to one another because of their interval of a perfect fifth, despite their linear distance in a standard chromatic scale, as depicted in Figure 3.2.

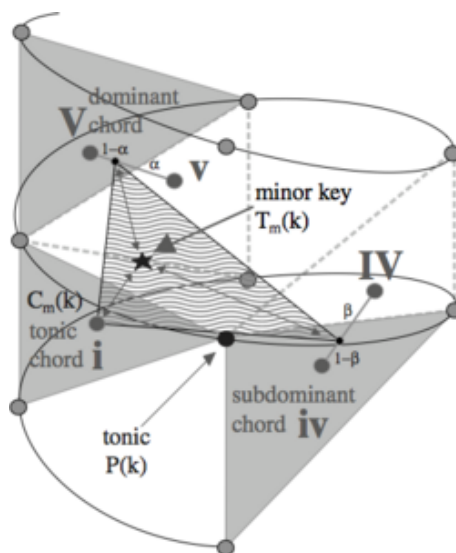


Figure 3.2: Spiral Array representation of a minor key as a center of effect (CE). The key is modeled as the CE of its primary triads (i , iv/IV , and V/v), which are themselves generated as the CEs of their constituent pitches on the outermost helix.

This hierarchical arrangement is beneficial for co-creative systems because it facilitates the real-time calculation of the center of effect. By treating musical passages as clusters of points within the array, an agent can mathematically determine the spatial centroid of a performance to

perform automated key-finding and harmonic tracking (Chew, 2014). A machine-learning partner can thus respond to individual musical notes and understand the changes in tonal centers during a human collaborator's improvisations.

Likewise, the Spiral Array connects symbolic analysis with the vector-based needs of modern AI. Because it maps discrete music theory concepts into a continuous, differentiable Euclidean space, it provides a topological ground truth for generative models (Chew and Chen, 2005). By working within this 3D space, a co-creative system helps keep its musical ideas understandable and relatable, even when performing complex interpolations across distant musical keys.

3.1.1.4 Geometric Orbifolds: Non-Euclidean Manifolds

Tymoczko (2006) revolutionized musical geometrical spaces by using the notion of orbifolds. These are mathematical spaces that loop back on themselves to account for musical symmetries. He demonstrates that if we treat a chord as a single entity regardless of its octave displacement or permutation, the resulting space of all possible chords is a non-Euclidean manifold. Depending on the number of voices and the specific constraints applied, these spaces often take the shape of high-dimensional tori or complex Möbius strips (Callender et al., 2008).

This model provides a clear proof of how Western harmony has evolved, showing that the most useful chords, those that sound good together and can be smoothly connected, tend to be found in specific regions of these mathematical shapes. However, as chords become more symmetrical, allowing for better voice-leading, they often lose their strong tonal center (Tymoczko, 2011).

For building co-creative interfaces, orbifolds offer a continuous map of harmonic algorithmic possibilities (Tymoczko, 2011). Unlike discrete symbolic grids, these spaces help find the shortest path (geodesic) between two points in a curved space, leading to the most efficient voice-leading transition between any two chords (Callender et al., 2008). By treating harmony as a path in a non-Euclidean space, the system can handle complex musical changes and minimal pitch variations with the same geometric precision that was once used for simple diatonic progressions (Pinto, 2012).

3.1.1.5 Fourier Phase Space: The DFT Perspective

A recent computational paradigm, led by Quinn (2006) and Amiot (2016), applies the Discrete Fourier Transform (DFT) to pitch-class sets to quantify their structural properties. By treating a set of notes as a periodic signal over the 12-tone chromatic circle, the DFT produces complex components derived from the vector sum of its component pitch classes. These coefficients make a chord's harmonic content manifest by characterizing its resemblance to various equal divisions of the octave, such as paradigmatic objects like the augmented triad (0, 4, 8) or diminished seventh chord (0, 3, 6, 9). For example, evenly spaced notes like those in a whole-tone scale create high coefficients in specific components. On the other hand, uneven patterns, such as those found in a major triad (0, 4, 7) or diatonic scale (0, 2, 4, 5, 7, 9, 11), have different coefficient profiles (Yust, 2017). In this framework, the coefficients act as vectors (with a magnitude and an angular position

(phase)) that serve as coordinates within a Fourier Phase Space. Unlike the Tonnetz or TPS, which lean on standard tonal functions, the Fourier perspective offers a mathematical method to understand a chord's tonal character by looking at its symmetry and interval structure (Yust, 2015b). Furthermore, recent theoretical work demonstrates a powerful convergence between these Fourier phase spaces and geometric voice-leading spaces. By interpreting continuous “crossfade paths” or glide paths through the Fourier space, this perspective can also represent contrapuntal voice leadings, mapping the smooth transition of notes from one harmony to another within a robust geometric context (Tymoczko and Yust, 2019).

This approach is particularly valuable in co-creative interfaces because it relates musical ideas with sound processing, allowing the model to analyze and adjust features of a hidden representation, like how diatonic or chromatic a musical piece is, without needing a detailed set of music theory rules (Lenssen and Needell, 2014). The DFT space also allows for a smooth transition between musical sets. Crucially, interpolating within the Fourier domain allows for smooth, continuous morphing between distinct musical sets, making it an invaluable tool for generative models to explore unconventional or microtonal pitch regions while maintaining structural consistency (Amiot, 2016).

3.1.1.6 The Tonal Interval Space (TIS)

The Tonal Interval Space (TIS), developed by Bernardes et al. (2016a), builds on the concept of Fourier Phase Space but takes a different approach by using high-dimensional vector spaces instead of traditional geometric shapes. In TIS, musical pitch collections are represented through twelve-dimensional (12D) Tonal Interval Vectors (TIVs) within a single continuous space. These vectors are calculated using the weighted DFT of 12-element, L1-normalized chroma vectors. This mathematical formulation makes it uniquely suited to analyze both audio and symbolic representations in real-time, facilitating the application of ML techniques.

TIS accommodates inputs from both symbolic data (e.g., MIDI) and complex audio signals by first converting them into chroma vectors, which distribute energy across chromatic pitch classes. In the resulting 12D Euclidean space, the spatial relationships between TIVs carry profound musical meaning. The Euclidean distance between two points indicates their perceptual relatedness, with shorter distances reflecting higher similarity, voice-leading parsimony, and common-tone logic (Bernardes et al., 2022). Additionally, TIS introduces a novel capability to measure musical consonance directly: the norm of a TIV (its Euclidean distance from the center of the space) acts as a consonance indicator, with points closer to the center representing dissonant configurations and points further out representing consonant ones (Bernardes et al., 2016a).

In this context, the direct application to audio signals bypasses the notoriously difficult problem of robust polyphonic note transcription, enabling systems to derive harmonic meaning without strictly relying on symbolic formats. This resilience to transcription errors makes the model exceptionally effective for real-time AI interactions and generative music systems. It creates a co-operative interface that can immediately respond to a performer's musical intensity while offering a unified operation chain to bridge different areas of creativity (Bernardes et al., 2016b)

Furthermore, TIS provides a stable framework for cross-domain mapping. Because the 12D vectors are continuous, they can be easily aligned with the latent spaces of neural networks or the parameter spaces of sound synthesizers. This continuity allows a co-creative interface to act as a topological mixer, where the user can navigate complex harmonic landscapes by moving through a simplified two or three-dimensional projection of the 12D space (Bernardes et al., 2018). By simplifying these complex relationships into an intuitive geometric interface, TIS helps performers adjust the tonal weight and harmonic brightness of a composition through physical, spatial interaction.

A practical application of this spatial interaction is demonstrated in the earGram 2.0 software, which utilizes the Tonal Interval Space to facilitate the creative re-purposing of musical audio. At the sound object level, earGram 2.0 provides navigable 2D spatial corpus visualizations where functional harmonic categories and key relations, such as those found in the circle of fifths, are expressed as geometric distances. To achieve this, the continuous 12D Tonal Interval Vectors are partitioned using QT-clustering and reduced to a simplified 2D visual representation based on the space's largest weighted coefficients, allowing users to visually guide the categorization process. By moving through this visualized harmonic space using an interface device like a mouse or joystick, a user can dynamically construct unit sequences at various density rates. Because reduced spatial distance within this model correlates with higher perceptual similarity, continuous multi-dimensional navigation provides the user with precise control over harmonic sequencing. This capability enables the generation of high-density textures of concurrent audio units, or discrete sound events, organized strictly by their calculated acoustic relatedness (Bernardes, 2020).

3.1.2 Rhythm and Metric Spaces: Temporal Topology

Rhythm and metric descriptor spaces reveal how musical events can be periodic and use timing expressively. Unlike traditional rhythmic notation, these spaces, shown in Table 3.2, depict rhythm as a smooth range of onset probabilities and timing relationships. This approach lets creative systems adjust the rhythm, shifting between a steady beat and the expressive timing of a live performer (Desain and Honing, 2003).

The literature maps this transition from discrete symbolic hierarchies to continuous generative spaces through several key models, which will be detailed in the next sections.

3.1.2.1 Metric Grids

Rooted in Lerdahl and Jackendoff (1982)'s Generative Theory of Tonal Music (GTTM), the metric grid acts as a psychological guide for how rhythms are perceived by organizing time into vertical layers of beats. Research by Desain and Honing (1994) shows that we do not hear rhythmic intervals by themselves. Instead, we compare them to an internal rhythm pattern. This process, known as categorical perception, allows us to recognize a note that comes slightly late as a delayed downbeat instead of hearing it as a new rhythmic value (Schulze, 1989).

Table 3.2: Typology of Rhythm and Metric Space Models: From Symbolic Hierarchies to Continuous Landscapes.

Model Category	Core Representation / Geometry & Key Authors
<i>Symbolic & Hierarchical Models (Discrete)</i>	
Metric Grids Lerdahl and Jackendoff (1982)	Hierarchical Dots: Vertical, layered representation of metric weight, distinguishing between structural beats and subdivisional pulses.
Circular Geometry Toussaint (2019)	Cyclic Polygons: Rhythmic cycles mapped as polygons on a circle to analyze mathematical symmetry and evenness.
<i>Statistical & Signal-Based Models (Descriptive)</i>	
IOI Distribution Desain and Honing (2003)	Statistical Histograms: Aggregated Inter-Onset Intervals (IOI) used to define a probabilistic profile of groove and pulse.
Periodicity Representations Peeters (2004)	Spectral Fluctuation: Signal-based representations (e.g., DTW) capturing modulation magnitude across frequency bands (e.g., Rhythm Patterns).
<i>Continuous & Generative Spaces (Latent/Morphable)</i>	
Metric Orbifolds Tymoczko (2011)	Manifolds: High-dimensional geometric spaces where all rhythmic partitions of a measure are treated as continuous points.
Probabilistic Distributions Kaplan et al. (2023)	Onset Density Functions: A continuous topography of onset probabilities; represents rhythm as a smooth landscape enabling micro-timing analysis.

Each vertical layer of the metric grid shows a different level of regularity in the beats: 1) the tactus is the basic beat that people often tap their foot to (level 0); 2) the subdivisions are quicker beats that break the tactus into smaller parts, like eighth or sixteenth notes (levels -1 , -2); and 3) macro-beats are slower beats that group tactus beats into measures and phrases (levels $+1$, $+2$).

For co-creative systems, embedding this hierarchy is vital because it enables the machine to calculate pulse salience, differentiating between structurally essential musical elements and fleeting ornamental notes (Parncutt, 1994), making it vital in real-time collaboration settings.

3.1.2.2 Circular Geometry

A significant shift in rhythmic representation occurs when the temporal sequence is mapped onto a circular topology. This approach, researched extensively by Toussaint (2019), treats a rhythmic

cycle (e.g., a single measure of 4/4 or a 12/8 African bell pattern) as a closed loop where the circumference represents the total duration of the period. By regarding onsets as points placed around a circle, we change rhythm from a straight line of events into a cycle of shapes (Toussaint, 2010). This approach helps researchers use methods from geometry and math to study musical time, uncovering important features that traditional Western notation often hides.

The power of circular geometry lies in its ability to quantify the shape of a rhythm through measures of symmetry, evenness, and balance (Toussaint, 2019). Maximal evenness patterns, where onsets are distributed as evenly as possible within a cyclic structure, often represented using circular diagrams and Euclidean algorithms (Gomez-Martin et al., 2008), help explain why specific rhythmic patterns, like Tresillo or Cinquillo, are common in many musical traditions around the world.

In a co-creative context, these geometric properties allow for computational evaluation of the stability of a rhythm. By calculating the centroid of the rhythmic polygon, a system can determine the balance of a pattern, providing a descriptor space where groove is defined not by subjective feeling, but by the mathematical displacement of vertices from a perfectly symmetrical state (Toussaint, 2019).

Furthermore, the circular model facilitates a unique form of rhythmic comparison through rotational equivalence and necklace theory (Toussaint, 2019). Because the representation is cyclic, a rhythm and all its possible rotations (or modes) are treated as a single family of patterns. This is particularly relevant for co-creative agents performing metric morphing, as it allows the system to smoothly interpolate between different phase-shifts of the same rhythm (Barton, 2020). Rather than seeing a shift in the one as a total rhythmic change, the system perceives it as a rotation in the circular space.

3.1.2.3 Inter-Onset Interval (IOI) Distribution

While geometric models dictate the nominal duration of notes, Inter-Onset Intervals (IOIs) capture the physical time between consecutive note onsets in an actual performance. By collecting these continuous intervals into statistical distributions, computational musicologists can map expressive timing variations onto discrete rhythmic notations (Cocharro et al., 2021).

Foundational theoretical models of quantization helped to establish how systems might bridge this gap (Longuet-Higgins, 1976). These foundational theoretical models often derive a discrete rhythmic interpretation by recursively subdividing a time interval (Longuet-Higgins, 1976). Furthermore, early studies on categorical rhythm perception demonstrated that listeners intrinsically group continuous temporal sequences into distinct categories, with boundaries that can shift depending on metrical context (Clarke, 1987a) and are open to perceptual learning (Schulze, 1989). Together, these probabilistic and categorical frameworks allow systems to generate a temporal fingerprint of a performance, identifying the underlying pulse even amidst complex variations.

The distribution of IOIs also highlights a complex relationship between global tempo and local expressive nuances. Earlier theories, such as the relational-invariance hypothesis explored by (Repp, 1994), argued that expressive timing scales approximately proportionally with the overall

tempo of a piece. However, subsequent analyses challenged this view, providing evidence that expressive timing variations do not simply scale proportionally with global tempo, but are instead inextricably linked to local musical structure (Desain and Honing, 1994). In human perception, while musicians produce these continuous timing deviations on a millisecond scale, listeners group them into discrete rhythmic categories, which then function as a reference point for appreciating the expressive deviations (Desain and Honing, 2003). For co-creative systems, analyzing these structural and expressive distributions allows machines to capture human nuances. It helps distinguish the relaxed, structural timing of a live performer from the rigid quantization of an electronic beat (Mor et al., 2020).

Furthermore, in the context of score-following systems, IOI sequences provide the foundational data for temporal alignment. Coupled with dynamic programming algorithms (Dannenberg, 1984) or more powerful methods, like Hidden Markov Models (HMMs) and hybrid graphical models (Orio and Schwarz, 2001; Raphael, 2006), IOI sequences can align polyphonic audio with musical scores. A common and highly effective continuous alignment method is Dynamic Time Warping (DTW), which finds the optimal match between a performer’s IOIs and the ideal score by non-linearly stretching the time axis to minimize the alignment cost (Müller, 2007). Because traditional DTW requires complete offline knowledge of both sequences, modern interactive systems utilize incremental online time-warping algorithms to align live performances in real time (Cancino-Chacón et al., 2018). By extracting these expressive parameters dynamically, co-creative systems can align precisely with the performer’s specific timing patterns, bridging the gap between a rigid timing grid and the fluidity of human rhythm (Echeveste et al., 2013).

3.1.2.4 Periodicity Representations

While IOI distributions effectively measure the discrete time between symbolic events, they often fail to capture the full polyphonic and continuous complexity of a musical arrangement. To address this issue, research in Music Information Retrieval (MIR) and speech prosody utilizes Periodicity Representations or modulation-theoretic approaches. Instead of counting the time between discrete notes, these models extract a continuous-valued onset-energy function (or amplitude envelope) from the audio and apply signal processing techniques such as the DFT or the Auto-Correlation Function (ACF). This yields a modulation spectrum (or beat spectrum) (Foote and Uchihashi, 2001), essentially treating rhythm as a form of low-frequency pitch defined by rhythm formants (i.e., frequency zones of high magnitude that correspond to regular rhythmic oscillations) (Gibbon, 2023).

The primary advantage of these periodicity descriptors in a latent space is their ability to represent hierarchical metric density across multiple frequency bands simultaneously. Just as low-frequency rhythm formants in speech correspond to slower word rates while higher frequencies align with rapid syllable rates, musical rhythm spectra distinctly encode different structural levels of a groove (Gibbon, 2023). A computational system can distinguish the low-frequency energy corresponding to the primary structural pulse from the high-frequency energy associated with rapid rhythmic subdivisions. By analyzing these multiband spectral fluctuations, a co-creative

agent can accurately extract and model the hierarchical metric density and rhythmic structure of a live performance (Peeters, 2011). If a performer plays with more intensity in their subdivisions, the high-frequency rhythmic energy increases, signaling the agent to make its own rhythmic accompaniment denser in response.

These descriptors enable the navigation of latent rhythmic spaces by providing high-dimensional, tempo-independent features that machine learning models can readily process (Peeters, 2011). Periodicity descriptors are exceptionally stable against small micro-timing noise, capturing the global feel of a groove far better than raw audio or discrete MIDI. Furthermore, by calculating a sequence of these spectra over short, overlapping windows to form a rhythm spectrogram, systems can extract rhythm formant trajectories. Building on foundational rhythmogram models (Todd and Brown, 1994), these trajectories can track the long-term, dynamic variations of rhythm over time (Gibbon, 2023).

3.1.2.5 Metric Orbifolds

While circular representations effectively model a single repeating rhythmic pattern, Metric Orbifolds conceptualize the universe of all possible rhythms of a given length within a unified geometric framework. Adapted from the continuous geometric spaces developed for harmony and voice-leading by Tymoczko (2006), Tymoczko (2011), and formally mathematically defined by Callender et al. (2008), this approach represents a rhythm of n events as a single coordinate point in an n -dimensional space. By mathematically abstracting away specific starting points, treating rhythms that are cyclic rotations of one another as equivalent, the fundamental multidimensional space is folded into a topologically complex quotient space, or orbifold. This geometric structure elegantly captures the symmetries and permutations inherent in cyclical musical rhythms, mapping equivalent patterns to the same location.

A critical advantage of the orbifold for computational musicology is its underlying continuity. In a traditional discrete metrical grid, transitioning from a duple subdivision to a triplet feel necessitates a categorical jump. In a Metric Orbifold, however, these two rhythmic feels exist as distinct coordinates on the same continuous surface, populated with every conceivable intermediate micro-timing between them. Such continuous transformations provide a rigorous mathematical definition of rhythmic distance, analogous to voice-leading size in pitch space (Callender, 2004): two points closer together on the orbifold are highly similar, requiring less aggregate movement of their temporal onsets to transform into one another. Consequently, complex structures like cross-rhythms (e.g., 3-against-2 or 5-against-4) are not treated as anomalous deviations from a grid, but rather as mathematically precise, symmetrically arranged coordinates within the space.

3.1.2.6 Probabilistic Distributions

The shift from rigid metrical grids to continuous models of temporal expectation gives rise to the concept of the probabilistic distributions of time. Foundational frameworks like Dynamic Attending Theory first modeled temporal expectations as continuous attentional pulses rather than static

points (Large and Jones, 1999). Building on this, recent cognitive frameworks grounded in Predictive Processing (Vuust et al., 2022), such as the Phase Inference from Point Process Event Timing (PIPET) model (Cannon, 2021), demonstrate that human rhythm perception can be simulated as a process of continuous Bayesian inference (Kaplan et al., 2023). In this probabilistic map, rhythm is represented by an expectation template formed by a summation of Gaussian distributions. The peaks (Gaussian means) indicate the anticipated phase of a musical event, corresponding to the underlying metrical grid. Meanwhile, the variance or width of these distributions captures the uncertainty of the expectation, effectively modeling expressive micro-timing, swing, and the specific probabilistic distributions that characterize laid-back or pushed live performances (Hosken, 2021).

This probabilistic view of rhythm is highly advantageous for co-creative systems because it treats metrical structure as a flexible, dynamic space rather than a sequence of static bins. By navigating rhythm as a probabilistic distribution, interactive agents can effectively manage real-time temporal uncertainty. If a performer introduces complex micro-timing deviations, the agent can dynamically adjust its internal confidence by updating the variance of its internal Gaussian phase estimates (Kaplan et al., 2023). Furthermore, translating discrete rhythmic patterns into continuous geometric spaces allows systems to interpolate smoothly between drastically different grooves. By weighting the features of surrounding patterns in a latent space (e.g., syncopation, frequency class, and density), algorithms can dynamically generate novel, hybrid rhythms that act as transitional states (Gómez-Marín et al., 2020). This represents a paradigm shift in rhythmic modeling: metrical structure is no longer conceptualized as a sequence of rigid, discrete events, but rather as a continuous, probabilistic distribution of temporal expectations. Within this framework, the computational distance between human and machine is quantified through the statistical divergence of their respective temporal distributions. This methodological approach facilitates continuous mathematical interpolation, enabling the system to dynamically synchronize and align its metric expectations with the human performer’s micro-timing variations.

3.1.3 Timbre Spaces: The Multidimensionality of Sound

While tonal and rhythmic elements often rely on clear mathematical or symbolic frameworks, timbre needs a different approach because it is more complex to computationally represent (Morrison, 2024). Timbre does not have a single physical measurement, so we describe it using different features that connect sound characteristics (e.g., sound spreading, change over time, and noise quality) with how people hear and experience sound (McAdams and Giordano, 2016).

These models, summarized in Table 3.3, have evolved from early Psychoacoustic Multidimensional Scaling (MDS) studies that identified the primary dimensions of instrument perception (Grey, 1977; Wessel, 1979), to MFCC-based spectral vectors that allow for the organization of vast sound corpora (Logan, 2000) and high-dimensional descriptor maps (Schwarz, 2007). Recently, new types of neural networks, like Self-Organizing Maps (SOMs), have learned to understand the layout of sound libraries. This helps create easier ways to explore different sounds (Cella, 2022). In the context of co-creative systems, these spaces act as a semantic mediator, allowing the human

Table 3.3: Typology of Timbre Space and Audio Descriptor Models.

Model Type & Key Authors	Core Representation / Geometry
Psychoacoustic MDS Grey (1977); Wessel (1979)	Euclidean Distance: Low-dimensional (2D or 3D) spaces derived from Multidimensional Scaling, where axes correspond to perceptual features such as brightness and attack.
MFCC Spaces Logan (2000)	Log-Spectral Space: Maps timbre texture using Mel-Frequency Cepstral Coefficients, providing a robust representation of spectral envelopes and speech-like qualities.
Descriptor Maps Schwarz (2007)	High-Dimensional CBCS: Utilizes standardized audio descriptors (e.g., MPEG-7) to organize large corpora into navigable clusters based on spectral and temporal features.
Self-Organizing Maps (SOM) Kohonen (1990); Cella (2022)	Neural Topographies: Employs unsupervised competitive learning to project high-dimensional audio features onto a 2D topological grid for intuitive spatial exploration.

and the machine to share a standard map of sonic textures and intensities (Schwarz et al., 2016). The next sections will detail each timbre space model.

3.1.3.1 Psychoacoustic Multidimensional Scaling (MDS)

The foundational work in timbre spatialization derives from non-metric MDS, popularized by Wessel (1979) and Grey (1977). From listeners' ratings of how similar pairs of instrument sounds were, researchers found that they could mathematically project those ratings into a two or three-dimensional Euclidean space (Krumhansl, 1989). In these spaces, the axes often correspond to specific psychoacoustic properties (i.e., x -axis to perceived brightness, y -axis to attack time, and other dimensions to roughness or spectral flux) (McAdams and Giordano, 2016).

A recurrent challenge in timbre modeling is linking the abstract dimensions of a perceptual space to quantifiable acoustic descriptors. Recent research has shown clear links between certain sound features and how we interpret music. These studies have improved our understanding of these relationships, highlighting that they can be complex and not always straightforward. The most important of these mappings are outlined in Table 3.4. While features like sharpness and spectral flux offer enough discriminative power to distinguish subtle instrument variations (McAdams et al., 1999) or support human echolocation (just-noticeable-difference near $0.04acum$, as noted in (Schenkman and Gidla, 2020)), other features like the spectral centroid are highly context-dependent and vary with harmonic density (Hermes et al., 2016). Consequently, perceptual qualities like roughness or fullness come from complex interactions across different areas.

This means that co-creative AI systems need to look beyond individual sound features. They should consider the many dimensions of how we perceive sound (Grey, 1977).

Table 3.4: Correlation between acoustic descriptors and timbre interpretation.

Perceptual Dimension	Acoustic Descriptors	Musical Interpretation
Spectral Envelope	Spectral centroid, spectral flux, spectral roll-off	“Brightness,” harmonic richness, spectral evolution over time.
Temporal Envelope	Log-attack time, temporal centroid, decrease	“Impulsiveness,” articulation, onset dynamics (e.g., plucked vs. bowed).
Harmonic Structure	Odd-to-even harmonic ratio, spectral kurtosis, inharmonicity	“Nasality,” “hollowness,” “brilliance,” or metallic textures.
Modulation	Spectral variation, FM parameters, roughness	Synthesis parameterization, tremolo/vibrato, timbre instability.

In a co-creative interface, this strategy offers a map of instrument timbre features, from which a performer can continuously navigate. Performers can uncover hybrid sounds, like a mixture between a “plucked string” and a “blown pipe”, by interpolating between known coordinates (Cae-tano and Rodet, 2011). Instead of just picking a sound from a set list, performers can explore a flexible sound space. This method sees timbre as a smooth shape, allowing the interface to adjust directly to specific sound details (Poirier-Quinot et al., 2017). As a result, the machine can create new sounds that don’t fit into traditional categories but remain perceptually coherent to the user (Hourdin et al., 1997).

Furthermore, MDS-based spaces facilitate a unique form of gestural mimicry in co-creative interfaces. Because the coordinates are grounded in human perception, an AI agent can analyze a performer’s input and respond by moving to an opposing or complementary region of the timbre space (Arfib et al., 2002). For example, if a musician makes their sound brighter, the system might automatically adjust to a darker sound to keep a balanced mix.

3.1.3.2 Mel-Frequency Cepstral Coefficients (MFCCs) Spaces

In music information retrieval and speech processing, Mel-Frequency Cepstral Coefficients (MFCCs) represent sound in a way that mimics how we hear. They were first developed for voice recognition. To create MFCCs, we analyze a sound signal using the Discrete Fourier Transform (DFT), which converts the sound’s power to the Mel scale, helping fit it to our pitch perception. Then, we apply a discrete cosine transform to the logarithm of the power values (Logan, 2000). This method produces a set of coefficients that represent the overall shape of the spectrum while mainly remaining unaffected by the fundamental frequency (f_0).

In co-creative settings, MFCC spaces work well for blending textures and vocals (Borsky et al., 2017). MFCCs give a clear and straightforward way to represent the identity of sounds, helping a system categorize sounds based on their general spectral features (Rawat et al., 2023). This ability enables different types of interactions, like connecting sounds across different forms. For

example, the timbre of a human performer's voice could manipulate the formant characteristics of a digital synthesizer by simply changing their vocal vowel shape. Thus, the MFCC space acts as the topological lookup table that translates these vocal gestures into synthesis control signals (Tzanetakis and Cook, 2002).

Furthermore, MFCC-based representations allow co-creative agents to perform automated segmentation and classification of a live performance. By tracking the trajectory of a performer through an MFCC space, an AI can identify shifts in playing technique, such as moving from *sul ponticello* to *ordinario* on a violin, and adapt its accompaniment logic accordingly (McKinney and Breebaart, 2003), making the MFCC space an essential tool for systems that require a high-level situational awareness of the sonic texture being produced by their human collaborators.

3.1.3.3 Descriptor Maps

As digital libraries expanded, the need for timbre mapping changed from categorizing a small number of instruments to organizing many thousands of individual sound grains. Schwarz (2003) tackled this issue by creating the descriptor map model. In this framework, every sound in a massive corpus is analyzed for a suite of standard audio descriptors, frequently using the MPEG-7 standard, including spectral flatness, roll-off, harmonicity, and temporal energy (Peeters, 2004). These sounds are then plotted as discrete points in a high-dimensional feature space. To make this space navigable for a human performer, dimensionality reduction techniques like MDS (Kruskal, 1964), Principal Component Analysis (PCA) (Jolliffe, 2002), T-distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton, 2008), or Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) are often used to project the cloud into a 2D or 3D visual interface.

For co-creative agents, descriptor maps enable high-speed real-time timbre matching. If a performer produces a loud, scratchy bowing gesture on a cello, the system can instantly calculate the descriptor coordinates of that input and trigger a granular sample with the nearest matching values (Caetano and Rodet, 2010). This allows the machine to follow the performer's sound changes, creating a shared audio experience that feels natural and responsive, rather than pre-planned.

Furthermore, these maps connect different sound environments (i.e., a live flute performance and a collection of industrial field recordings). By navigating the spectral brightness or noisiness of the flute, the performer effectively plays the industrial sounds, using the descriptor map as a relational map that preserves the gestural intent across disparate sonic domains (Schwarz et al., 2006).

3.1.3.4 Self-Organizing Maps (SOM)

A more advanced computational approach to timbre organization utilizes Self-Organizing Maps (SOM), an unsupervised neural network technique developed by Kohonen (1990). Unlike MDS, PCA, or t-SNE, which are linear dimensionality reduction techniques, a SOM is trained to learn

the underlying topology of a high-dimensional dataset, such as a library of several thousand orchestral samples, and project it onto a lower-dimensional grid (typically 2D). This process keeps the neighborhood relationships of the original data intact. It makes sure that sounds that are similar to each other are placed next to one another on the map.

In the context of modern computer-aided orchestration, frameworks such as Orchidea (Cella, 2022) utilize these neural topographies to manage the vast search space of orchestral combinations. For a co-creative interface, SOMs provide an intuitive navigational surface where a user can explore a sound world through spatial gestures. Instead of using complex controls like brightness or attack, the performer interacts with a learned topography of sound. By moving a cursor or hand across the grid, the performer can create smooth transitions between different sounds, such as metallic tones and wooden percussion, in a continuous way (Cella, 2022).

This paradigm is extended from static timbre mapping to dynamic musical behavior in the MASOM (Musical Agent based on Self-Organizing Maps) framework (Tatar and Pasquier, 2017). MASOM utilizes SOMs to organize musical units into a latent space that represents both acoustic features and temporal contexts. By using this learned layout, the agent can improvise in real-time based on style. It can choose musical elements that sound good together and match the style of the ongoing performance. In this sense, the SOM acts as a bridge between low-level signal processing and high-level creative agency, allowing the system to respond to a human performer with stylistically coherent gestures derived from its trained corpus (Tatar and Pasquier, 2017).

Furthermore, SOMs allow for interpolation between highly disparate sounds. Because the space is organized based on statistical similarity, the empty cells between two learned sound clusters represent mathematical interpolations of those timbres (Toivainen, 1997). This enables a co-creative agent to propose new sounds that exist in the gaps of the training data, facilitating a form of collaborative sound design where the human navigates the map and the machine synthesizes the intermediate results (Thelle and Pasquier, 2021; Thelle, 2022).

3.2 The Shift to Latent Representations

The way we understand and use descriptor spaces has changed with the rise of deep generative modeling. Traditionally, audio features like spectral centroid or zero-crossing rate were defined by humans in systems like CBCS. Currently, more people are using latent spaces in their work. These are compressed representations of high-dimensional data learned directly by neural networks during the process of dimensionality reduction and reconstruction. In this context, the machine does not follow a predefined musical taxonomy. Instead, it discovers an internal, statistical organization of sound (Roberts et al., 2018).

3.2.1 From Prescriptive Features to Emergent Topologies

A neural latent space uses connected axes to intrinsically represent sound qualities. Unlike traditional descriptor spaces, which are derived either from mathematical formulations or cognitive models, this system learns a continuous shape that captures the main aspects of a group of sounds.

This change affects how artists create music. Instead of picking sounds from a list of labels, they explore a space of possibilities (Vigliensoni et al., 2020).

The main benefit of the latent space is its ability to connect different sounds smoothly. Because the space is continuous and shaped by the training data, moving between two points results in a gradual, clear change in the audio. This process, called interpolation or timbre morphing, lets one change sounds in creative ways. It goes beyond traditional sound synthesis and blurs the lines between composing and performing (Akten et al., 2020).

3.2.2 Latent Space Deep Models

In the landscape of generative deep learning, several architectures are designed to compress complex musical data into lower-dimensional latent spaces. Variational autoencoders (VAEs) (Kingma and Welling, 2014) are the primary models used for real-time interaction, as they provide a smooth, bi-directional mapping between raw data and a navigable manifold. This system learns to convert complex audio signals into a simpler format, known as a bottleneck, where important sound features are kept as latent vectors (Roberts et al., 2018). A VAE, different from a regular Autoencoder, organizes its encoded information in a way that follows a specific shape, usually resembling a Gaussian distribution. This setup helps ensure that the space is filled and continuous, meaning every point in the space represents a possible sound, even if that sound was not part of the original training data. For the co-creative artist, the VAE provides the connective tissue between known musical states.

Generative Adversarial Networks (GANs) can also produce latent spaces, though they often require additional inversion steps to map existing sounds into the space (Dong et al., 2018), while Transformers typically utilize a discrete latent space through Vector Quantization (VQ-VAE) to handle long-term structural sequences (Huang et al., 2019b). More recently, Diffusion Models have emerged as a high-fidelity alternative for timbre latent spaces, though their iterative denoising process often imposes significant computational latency compared to the instant, single-pass inference of autoencoders (Mariani et al., 2024). Thus, we will not expand on Diffusion Models in this thesis.

3.2.2.1 Tonal and Harmonic Latent Spaces

Tonal latent spaces must consider the complex relationships between pitches. Proximity is determined not only by frequency, but also by harmonic function. Models like MusicVAE (Roberts et al., 2018) and DeepBach (Huang et al., 2019a) see tonal space as a network of connected paths. The most significant advancement in this area is the creation of disentangled tonal spaces for separating melody from harmony, allowing for the independent adjustment of each (Bryan-Kinns et al., 2024).

In polyphonic models, such as the PIANOTREE VAE (Wang et al., 2020b), the latent space is organized into a hierarchy where the higher levels represent the harmonic progression (the chords) and the lower levels represent the melodic ornamentation. By exploring specific latent dimensions,

a creative agent can keep the basic identity of a song while changing the melody. This way, blending two songs leads to a smooth musical change that stays true to the main key, rather than creating a messy group of notes. (Dieguez and Soo, 2020).

While not a VAE in the traditional sense, Coconet (the model behind the Bach Doodle) utilizes a latent representation of counterpoint. It treats tonal space as a completion problem, where the machine understands the latent relationships between four-part voices. This allows a performer to provide a single melodic trajectory, while the system infills the harmonic landscape based on the learned logic of common-practice tonality (Huang et al., 2019a).

Modern research into VAEs informed by music theory attempts to force the latent space to adhere to known geometric structures like the Tonnetz. In these models, the distance between two latent vectors z_1 and z_2 is mathematically mapped to the harmonic distance between keys (Bryan-Kinns et al., 2021). For a co-creative agent, this means that a morph from C Major to G Major follows a shorter path in the latent space than a morph to F# Major. This alignment of neural statistical logic with human musical logic creates a more intuitive interface for live performers, as the latent space effectively understands the emotional and structural gravity of different keys (Wang et al., 2020a).

Beyond VAE-based manifolds, other architectures like MuseGAN and Music Transformer explore structural logic through different computational lenses. MuseGAN utilizes a GAN framework to model multi-track polyphony, emphasizing the relational logic between different instruments. Its latent space is often organized around the simultaneous interaction of tracks, though it lacks the smooth, bi-directional encoding found in VAEs, making it more suited for autonomous generation than for real-time gestural tuning (Dong et al., 2018).

In contrast, the Music Transformer leverages self-attention mechanisms to handle long-range dependencies and complex formal structures. While its primary strength lies in maintaining structural coherence over minutes rather than seconds, it operates within a discrete latent space. This makes it highly effective for thematic development and structural logic, as the model can remember a motive from the beginning of a piece and transform it later (Huang et al., 2019b).

3.2.2.2 Rhythmic Latent Spaces

In the domain of rhythm, the challenge for a VAE is to maintain long-term structural coherence across a measure. MusicVAE addresses this by using a hierarchical decoder (Roberts et al., 2018). Instead of generating a sequence of notes all at once, it generates conductor embeddings that then dictate the patterns for smaller sub-segments.

The Groove Variational Autoencoder (GrooVAE) architecture provides a technical solution to the problem of separating musical structure from human expression. GrooVAE represents drum patterns in a different way than traditional models. Instead of using a simple yes or no for whether a note plays, it breaks down each step into three parts: the onset (if a note plays), the velocity (how hard it is hit), and the micro-timing offset (how much it varies from the expected timing in milliseconds). This architectural design enables the latent space to function as a structured multidimensional manifold where two-dimensional coordinate axes (i.e., x and y) explicitly encode

the fundamental metric framework of the rhythm. In contrast, additional latent dimensions encode the expressiveness of the performance (Gillick et al., 2019).

The primary innovation of GrooVAE is its capacity for style transfer within the rhythmic domain. Because the model disentangles the score from the performance, a co-creative agent can take a quantized, metronomic MIDI pattern (the grid) and re-render it through a specific point in the latent space that contains the micro-timing profile of a professional drummer (Gillick et al., 2019). This is a practical example of metric morphing. The system does not just change the timing randomly; it follows a learned pattern that makes it seem more human-like.

Furthermore, this model facilitates a human-in-the-loop interaction where the latent space acts as a mediating descriptor space. A performer can use a single slider to move between a mechanical 16th-note hi-hat pattern and a laid-back swing feel (Gillick et al., 2019). Mathematically, this is achieved by shifting the latent vector towards regions of the manifold characterized by high variance in the offset dimension. By combining this with real-time synthesis like RAVE, a co-creative system can change both the rhythmic landscape and the timbre texture. This creates a musical work that responds directly and evolves alongside the human performer.

3.2.2.3 Timbre Latent Spaces

Most variations of VAEs focus on working with MIDI data, which represents music in a symbolic format. However, when it comes to audio, things get more complex. We need to understand and recreate the rich details of sound. In the past, early models used Conditional Variational Autoencoders (CVAEs) to transform frames of audio data, analyzed using Short-Time Fourier Transform (STFT), into simpler forms called latent vectors. A foundational example is the NSynth (Neural Synthesizer) project (Engel et al., 2017), which used the WaveNet Autoencoder to learn a latent representation of instrument sounds. While NSynth demonstrated the power of timbre morphing (e.g., interpolating between the latent vectors of a flute and a bass to create a hybrid instrument), the architecture was computationally expensive and lacked the real-time capabilities required for improvisatory performance (Engel et al., 2017).

The search for more efficient ways to represent data led to the use of VQ-VAE. This method breaks the latent space into a collection of acoustic atoms. Models like GANSynth tried VQ-VAEs to accurately connect data modeling with sound music creation. GANSynth utilized GANs to produce high-resolution audio from a latent space that explicitly separated pitch from other timbre attributes. This was a critical step toward disentangled control, but the reliance on frequency-domain processing often introduced phase smearing, which degraded the temporal resolution and acoustic fidelity of the synthesized output (Engel et al., 2019).

A recent approach to this task was the introduction of Singing Voice Conversion (SVC) models and the Differentiable Digital Signal Processing (DDSP) framework. DDSP combines the clear understanding of traditional sound creation methods, like using oscillators and filters, with the learning abilities of neural networks (Engel et al., 2020). This creates an efficient space where different features relate to physical aspects such as loudness and pitch. However, while DDSP is

effective, it is still limited by the specific sound creation model it uses. This limits its ability to handle complex sounds that don't follow simple patterns.

A significant milestone in the shift toward latent navigation is the RAVE (Realtime Audio Variational Autoencoder) architecture, which represents the synthesis of these various approaches. By using raw waveforms instead of spectral representations and applying a multiband decomposition strategy, RAVE achieves very low latency of less than a millisecond while maintaining rich sound quality. (Caillon and Esling, 2021).

The primary technical innovation of RAVE lies in its departure from the Auto-regressive (one sample at a time) and STFT-based (spectral frame) paradigms. Auto-regressive models like WaveNet are naturally slow because each sample relies on the prior one (Oord et al., 2016). In contrast, spectral models often suffer from phasiness and latency due to the windowing requirements of the Fourier transform (Engel et al., 2019).

RAVE bypasses these limitations through a multiband decomposition strategy. The model breaks the input signal into several frequency bands using a filter bank. This lets the convolutional layers work at a lower sampling rate while still capturing all frequency ranges. This “divide and conquer” method reduces the amount of processing needed, allowing for responses in less than a millisecond on normal consumer hardware. As a result, it creates a level of control that was not possible with deep generative models before. (Fels et al., 2002).

Unlike earlier lookup table models, RAVE's latent space is regularized as a continuous manifold. Because the model is trained as a VAE, the space between two sound points is mathematically smooth (Kingma and Welling, 2014). This allows the performer to traverse the manifold using gestural sensors (such as XY pads or accelerometers) to systematically navigate the multidimensional space, facilitating continuous acoustic interpolations between distinct timbre textures. It also allows the performer to manipulate the latent space. Since the decoder is streaming, any change to the latent coordinate is reflected immediately in the output audio (Caillon and Esling, 2021), enabling the performer to use the latent space as a dynamic filter rather than a static sound source.

In the context of digital organology, RAVE shifts the definition of the instrument. In traditional synthesis or CBCS, the instrument's body is the corpus or the oscillator bank (Schwarz, 2007). In RAVE, the instrument's body is the latent space. The performer does not interact with the sound directly but with the representation of the sound (Magnusson, 2009).

This establishes a mediating environment where the computational model applies the structural constraints of the learned corpus, such as maintaining the spectral density of the training data. Concurrently, the human performer supplies the physical input and directional control (Bevilacqua et al., 2017). This implementation of the mutable musical work enables the performer to generate novel, interpolated sounds that were not present in the original recordings but are statistically inferred from the continuous latent space.

3.2.3 The Latent Space as a Performative Territory

This research focuses on the VAE architecture because it is designed for effective navigation and structural interpolation. Unlike GANs, which can be messy and inconsistent, VAEs use a specific method called the KL Divergence to keep the latent space smooth and continuous. This smoothness prevents gaps in the space, allowing a performer to transition between different musical states, like varying harmonic functions or saxophone sounds, without producing unwanted noise. Additionally, the VAE’s encoder accurately maps an existing collection of data into the space. This combination of fast response, two-way mapping, and consistent structure makes the VAE the best choice for a mutable interface that provides reliable, instrument-like feedback.

By using VAE latent representations instead of descriptions written by humans, the musical work transforms into a space for performance. It becomes a complex area full of different possibilities, where changing structures is a key feature of the medium. As illustrated in Table 3.5, this shift marks a change from strict database querying to a more flexible way of navigating using interpolation and trajectory.

Table 3.5: Comparison Between Traditional Descriptor Spaces and Neural Latent Spaces.

Attribute	Traditional Descriptor Spaces)	Neural Latent Spaces (VAEs)
Feature Extraction	Human-defined (Spectral, Temporal)	Machine-learned (Emergent, Entangled)
Logic	Prescriptive / Symbolic	Descriptive / Connectionist
Navigation	Database Querying	Topological Interpolation
Performative Role	Selection and Triggering	Trajectory and Morphing

The shift to connectionist logic and topological navigation means that performers are moving from choosing specific units to navigating continuous spaces. This change forms the basis for what this thesis introduces as a unified latent territory. Within this proposed framework, the isolated successes of models like RAVE, GrooVAE, and MusicVAE are conceptually synthesized into a singular, performative landscape. By breaking down the music signal into three parts, timbre, rhythm, and pitch, this approach creates a flexible space where the fundamental elements of music serve as dynamically manipulable variables within a continuous, bidirectional feedback loop between the human performer and the computational system.

In this proposed framework, the mutable musical work emerges from the simultaneous and selective navigation of these three dimensions. This synthesis helps us envision a context for performance where the basic elements of music, which are usually seen as fixed instructions, are understood as things that can change continuously. For example, a performer might keep a steady rhythm while exploring different musical notes. This allows them to transition smoothly between various harmonic areas. Similarly, real-time controllers could be used to manipulate RAVE’s timbre space while maintaining the structural integrity and metric coherence provided by the rhythmic and tonal models. Finally, because these spaces are probabilistic, the system

can adapt to a performer’s energy level by blurring or sharpening the peaks within the rhythmic or tonal landscapes, effectively mediating between the rigid ideal of the score and the embodied reality of the performance.

Ultimately, this proposal presents the latent space as the best way to enhance spaces. It does not just store audio data, but also creates a detailed map that supports and improves our understanding of music. By transforming the discrete categories of Western music theory into a continuous, differentiable topology, the unified latent territory represents a shift toward topological composition. In this paradigm, the musical work is not the audio output itself, but the design of the latent space and the specific trajectories a performer takes through it. This collapses the distinction between composition and performance, providing a continuous map where algorithmic precision and human micro-timing converge.

3.3 Corpus-Based Concatenative Synthesis (CBCS) as a Space Enhancer

The move from timbre descriptor spaces, especially the self-organized topographies of MASOM (Tatar and Pasquier, 2017) and latent spaces, to the audio output of CBCS marks a change from modeling music structure to recreating sound. The SOM outlines different musical behaviors, while CBCS is the tool that brings these behaviors to life by carefully choosing and assembling audio grains (Schwarz et al., 2016). In this setup, the SOM acts as a behavioral filter for the audio collection. Instead of working with raw sound descriptors, the system operates in a learned space where each point helps in selecting audio units. In the field of digital instrument design, CBCS effectively uses descriptor spaces, turning a static audio collection into an interactive, multidimensional space (Schwarz, 2012).

3.3.1 The Topological Framework of CBCS

Based on Schwarz (2003)’s idea of descriptor maps, CBCS systems construct new sonic material by selecting and concatenating small sound units (grains) from a pre-analyzed corpus, guided by multidimensional descriptors that encode perceptual characteristics (Schwarz et al., 2006). These descriptors include features encompassing spectral, temporal, and gestural information. Together, they create a descriptor space that helps the system find sounds matching specific user-defined qualities (Schwarz et al., 2016). This means CBCS enhances the representation of the musical collection. It allows for detailed exploration and mixing of sounds while keeping them sounding coherent (Schwarz et al., 2007).

CBCS permits users to explore sounds by showing them as points in a multidimensional space. This setup makes it easy to understand different qualities of sounds, including their character and structure. The closer the points are to each other, the more similar the sounds are. Additionally, the layout helps reveal connections between different expressive qualities like brightness, roughness, or articulation (Schwarz and Schnell, 2009). Because the space is continuous, composers and

performers can traverse it interactively, mosaicing target audio by finding the k -nearest neighbors in the corpus that match desired descriptor coordinates. This approach looks at instruments not as physical objects but as the space where sound characteristics exist. The main job of the performer is to explore these complex sound areas (Schwarz, 2012).

3.3.2 Navigational Strategies in Concatenative Sound Synthesis

The navigation of CBCS spaces has evolved through distinct paradigms, continually attempting to bridge the gap between high-dimensional data and musical intent. Initially, manual and interactive navigation allowed performers to engage directly with descriptor spaces via visual manipulators. In systems such as CataRT (Schwarz et al., 2006), Nuvolet (Comajuncosas et al., 2011), and vr-Grains (Zappi et al., 2012), this point-and-click approach enables the immediate triggering of timbre neighborhoods, though it requires significant spatial memorization of the corpus layout. The paradigm subsequently shifted toward target-based navigation, or mosaicing, where the system automatically selects units matching the descriptor profile of an external target audio stream, facilitating real-time timbre mimicry.

To facilitate more coherent structural discovery, several systems adopt clustering strategies to organize the corpus into manageable, navigable topologies. For instance, the earGram framework implements algorithms such as k -means, quality-threshold (QT) clustering (Heyer et al., 1999), and density-based clustering (DBSCAN) (Ester et al., 1996) to group sound objects with shared perceptual characteristics, thereby dividing the corpus into representative, interactive sub-spaces (Bernardes et al., 2013). By exposing the higher-level organization of the audio sources, these clustering algorithms allow performers to navigate massive databases intuitively based on structural similarity. A parallel approach to this clustering logic is utilized in Google’s The Infinite Drum Machine,¹ which applies machine learning (specifically t-SNE) to organize thousands of everyday sounds into a navigable, clustered map based strictly on audio similarity, promoting highly intuitive, exploration-based navigation without requiring manual tagging (Google Creative Lab et al., 2016).

Building upon these clustered spaces, navigation can be further automated through generative sequencing models. In earGram, computer-aided algorithmic composition (CAAC) strategies replace manual point-and-click traversal by automating the selection procedures. earGram utilizes probabilistic models (i.e., Markov chains) to generate continuous, non-repeating sonic structures that emulate the temporal morphologies of the original audio (Bernardes et al., 2013). This is particularly effective for macro-structural navigation, such as in its soundscapeMap and infiniteMode environments, where the vertical and horizontal aggregation of units is driven by psychoacoustic rules rather than direct gestural control.

Similarly, sequence-modeling algorithms such as the Factor Oracle (FO) and its advanced extensions, namely the Audio Oracle (AO) and Variable Markov Oracle (VMO), automate the recombination of audio fragments by rigorously parsing the temporal morphology of the audio

¹The open-source repository for this project is available at <https://github.com/googlecreativelab/aiexperiments-drum-machine>, last accessed January 29, 2026.

source. In recent iterations of earGram (earGram 2.0), this automation is achieved by adopting the FO automaton to translate audio units into a discrete and finite alphabet of symbols derived from multiple perceptual viewpoints, ensuring that units with high perceptual affinity share the same symbol (Bernardes, 2020). By mapping these symbols through a network of forward factor links and backward suffix links, the FO identifies repeating sub-sequences and determines optimal splice points (Bernardes, 2020). Ultimately, traversing this automaton allows earGram to generate new unit sequences that guarantee the formal structural retention and sonic identity of the original material without simply looping (Bernardes, 2020).

Building upon this finite-state foundation, the VMO elevates the architecture to process multivariate time series for both symbolic and audio representations, optimizing its symbolization through an Information Rate (IR) metric (Wang et al., 2016). Crucially, this optimization empowers the VMO to perform guided improvisation, utilizing the live performer’s tonal chroma or rhythmic beat phase as a continuous query to drive corpus navigation (Wang et al., 2016). This query-matching functionality seamlessly addresses performative challenges like the “cadence” and “queueing” problems, ensuring the machine’s generative output dynamically synchronizes with human musical intent (Wang et al., 2016). To further elevate the musical output, the VMO implements structured improvisation via a heuristic switching mechanism that compares the longest repeated suffixes (lrs) of the generated phrases (Wang et al., 2016). By strategically oscillating between tracking the human input and exploring its own internal generative history, the algorithm successfully balances musical expectation with surprise, autonomously constructing robust macro-structural forms without requiring manual descriptor-based navigation (Wang et al., 2016).

State-of-the-art architectures, such as Somax2 (Borg et al., 2022; Fiorini and Malt, 2023), expand this automated sequencing into fully reactive co-improvisation environments (Assayag et al., 2022; Brochec et al., 2025). Rather than navigating linearly, Somax2 reorganizes the corpus into multidimensional structures based on textural and harmonic similarities, creating what Malt and Sluchin (2025) terms “listening topologies.” To operationalize this, the framework employs a dynamic cognitive memory architecture, specifically utilizing a shift/decay/merge mechanism, that probabilistically matches the performer’s live input to the system’s learned latent space (Assayag et al., 2022; Brochec et al., 2025). Ultimately, this mechanism transforms corpus navigation from a rigid sequence of discrete selections into a continuous, fuzzy pattern recognition of the unfolding musical context (Assayag et al., 2022).

Despite these advancements, as control systems grow in complexity, the challenge of mapping intimacy becomes more acute (Fels, 2004). While continuous gestural control provides a highly embodied interface, it often introduces mapping opacity, overwhelming the performer with configuration parameters and increasing cognitive load (Hunt et al., 2000; Best et al., 2018). To resolve this barrier to control intimacy, recent developments in Somax2 replace low-level descriptor mapping with real-time semantic classification, such as the recognition of Instrumental Playing Techniques (IPTs) (Brochec et al., 2025). By navigating the corpus through high-level morphological gestures, such as specific flute extended techniques, rather than discrete audio descriptors, the system dynamically aligns its generative output with the performer’s artistic intent (Brochec

et al., 2025). This paradigm shift from manual parameter control to intent-aware latent navigation significantly reduces cognitive load, allowing the performer to intuitively engage in what the literature defines as “playing with music,” fostering a genuine, unburdened co-creative dialogue (Malt and Sluchin, 2025).

However, this transition toward latent space navigation introduces a new challenge: the inherent opacity of deep neural networks. While delegating complex feature mapping to intent-aware algorithms effectively reduces the performer’s cognitive load, it risks transforming the co-creative system from a comprehensible tool into an unpredictable black box, fundamentally obscuring the system’s structural logic (Bryan-Kinns et al., 2024). As recent literature emphasizes, if musicians cannot clearly perceive how their gestural actions influence the model’s generative output, this lack of clarity inevitably undermines trust and restricts their overall creative involvement (Abdul et al., 2018). Consequently, to ensure these advanced navigational paradigms function as reliable collaborative partners rather than obscure mechanisms, it is essential to establish transparency and interpretability within these complex neural architectures.

3.4 Explainable AI in Creative Music Systems

As ML and DL become more important in creative practices, issues like transparency, understanding, and control are becoming more urgent. The complex nature of deep neural networks makes them powerful but also hard to understand. This lack of clarity can make it difficult for performers to see how their actions influence the model, turning it from a helpful partner into an unpredictable black box. In this situation, Explainable AI (XAI) is essential for making these complex systems easier to understand (Abdul et al., 2018).

According to Bryan-Kinns et al. (2024), it’s important for generative music systems to be understandable to build trust and encourage user involvement. If users don’t grasp how the machine works, they feel less in control and less connected to their creative expression. For effective collaboration in music creation, composers need to know how the system processes input and creates sound. This clarity helps maintain a meaningful interaction rather than just random outcomes (Fiebrink, 2011).

3.4.1 Levels of Explainability in Music

Recent studies Bryan-Kinns et al. (2024) on how humans interact with AI show that XAI in complex music settings can be understood on three main levels:

1. **Representational Level** focuses on making the model’s internal bottleneck visible. Techniques like t-SNE or UMAP create visual maps from machine-learned data, helping users understand how sound and movement features are represented.

2. **Interactional Level** concerns the real-time feedback loop. Interfaces that show how the system makes decisions or how certain it is help users adjust their input. As noted by [Bryan-Kinns et al. \(2024\)](#), this kind of transparency is important for creating a shared mental model between humans and machines.
3. **Conceptual Level** integrates interpretability as an aesthetic property. Here, the model's structure, such as the hierarchical organization of MusicVAE, becomes an explicit part of the dialogue. The performer engages with the machine as an intelligible partner whose style is a navigable territory of learned probability rather than an accidental output.

3.4.2 Latent Disentanglement and Semantic Control

One major challenge in latent navigation is that machine axes often lack clear musical meaning. A single dimension of a VAE can affect both pitch and brightness at the same time. This creates a mapping opacity that makes it hard to perform intentionally. Recent research on latent disentanglement works to solve this problem by separating different creative factors into distinct, independent dimensions. ([Higgins et al., 2017](#); [Burgess et al., 2018](#)).

Disentanglement bridges the gap between signal processing and musical concepts. In the tonal domain, it allows for the manipulation of functional concepts (i.e., tonic vs. dominant) without altering melodic contour ([Prang and Esling, 2020](#); [Pati and Lerch, 2021b](#); [Bryan-Kinns et al., 2024](#)). In the timbre domain, some models, like those by [Limberg and Zhang \(2024\)](#) and [Schwär et al. \(2022\)](#), map latent dimensions to perceptual timbre descriptors. Other models also attempt to disentangle pitch from timbre ([Yin-Jyun Luo et al., 2020](#); [Tanaka et al., 2021](#)).

This disentanglement is often achieved by regularizing the latent space via auxiliary losses to force the manifold to inherit the topology of known audio descriptors ([Pati and Lerch, 2021a](#)). For instance, [Esling et al. \(2018\)](#) uses targeted regularization to ensure specific axes control single perceptual attributes. By decoupling these features, the space changes from a confusing black box into a clear and understandable object. This allows for easier navigation, making it possible to move through the space meaningfully. Complex sound changes are simplified into a single, deliberate action ([Bryan-Kinns, 2024](#)).

3.4.3 Transparency as a Creative Catalyst

Bringing explainability into neural frameworks emphasizes the importance of understanding and communication in human-machine relationships. Transparency does not mean oversimplifying things. Instead, it creates a chance for an ongoing conversation. As [Bryan-Kinns et al. \(2024\)](#) points out, “seeing inside” the model lets the creator interact with the hidden aspects not as a barrier but as a helpful partner.

Shared ideas and feedback allow both human and machine agents to develop their musical identities together. These ideas form the basis for the research that follows, focusing on exploring hidden spaces, sound quality, and movement in cooperative electroacoustic practice.

3.5 Summary: Conceptual and Theoretical Framework of the Mutable Musical Work

Chapter 2 introduces the idea of descriptor spaces, which organize musical information across three key areas: 1) tonal spaces, which map notes, chords, and intervals' relationships; 2) rhythm and metric spaces, capturing timing and rhythm; and 3) timbre spaces, for classifying sounds based on their texture and quality.

The final sections of this literature review focus on balancing the high-dimensional complexity of modern musical models with the critical need for human interpretability. Current advancements, specifically neural latent representations, facilitate the continuous and fluid manipulation of foundational musical elements like pitch, rhythm, and timbre. However, to prevent these powerful generative systems from operating as opaque black boxes, it is essential to employ XAI methodologies. By utilizing clear visualizations and structural explanations, these complex topological spaces are rendered intelligible, ensuring the system functions as a transparent, collaborative partner in the co-creative process.

Ultimately, Chapters 2 and 3 emphasize that the mutable musical work is an evolving process rather than a fixed creation, shaped by the interactions between human performers and machine systems. The vision is for a music-making process that highlights collaboration and adaptability, with systems that enhance the performer's experience.

Chapter 4

Research Methodology and Design

While the theoretical frameworks explored in Chapter 2 envision a future of seamless human-machine collaboration, and Chapter 3 establishes the high-dimensional latent spaces required to map this interaction, the practical realization of such systems is hindered by a lack of integrated methodologies. Existing research often bifurcates into either purely technical evaluations of computational models or subjective accounts of artistic practice. To bridge the gap between abstract mathematical potential and real-world compositional and performative affordance, this chapter outlines a mixed-methods research design. By combining experimental, hypothesis-driven research assessed with quantitative methods and practice-based reflection, this methodology investigates the development of mutable interfaces that are both computationally robust and artistically viable.

The remainder of this chapter is structured as follows. Section 4.1 defines the problem statement, identifying the critical gaps between the mathematical potential of high-dimensional latent spaces and their compositional and performative affordances. Section 4.2 outlines the research questions and sub-questions that guide this investigation, specifically focusing on the ontology of the mutable work and the performative dynamics of contextual navigation. Section 4.3 details the operationalization of the research design, explaining the mixed-methods approach that integrates rigorous experimental validation with practice-based performative prototyping in an iterative cycle. Section 4.4 presents the implementation roadmap, organizing the study into a two-phase design covering symbolic tonal representation and expressive timbre navigation. Finally, Section 4.5 provides a summary of the research methodology.

4.1 Problem Statement: Bridging the Gap between Manifold and Performance

The central challenge of this research lies in the transition from the mathematical potential of high-dimensional manifolds to their compositional and performative affordance. This thesis identifies three critical friction points:

1. **The Opacity of Latent Agency vs. Performative Intent:** High-dimensional latent spaces offer immense generative power but suffer from a lack of interpretability. This research investigates how Explainable AI (XAI) can bridge the gap between model complexity and artistic agency.
2. **The Conflict Between Formal Logic and Sub-Symbolic Fluidity:** There is a methodological disconnect between symbolic music theory (rules/notation) and sub-symbolic neural audio (raw signal). This work seeks to anchor fluid neural landscapes within the formal constraints of musical logic.
3. **The Complexity of Embodied Mapping:** Mapping musical motion to high-dimensional audio parameters remains a significant challenge. This thesis addresses the problem of designing a contextually bounded environment where motion-informed data acts as a continuous bridge between the performer's intent and the system's topological boundaries.

4.2 Research Questions and Sub-Questions

To investigate the transition of the musical work from a fixed artifact to a mutable interface, the two primary research questions, RQ1 (ontology) and RQ2 (practical applications and contextual navigation), defined in Chapter 1, are extended into specific technical and performative sub-questions (SQ). This refinement ensures that the abstract concepts of agency and topology are operationalized through the experimental and practice-based tracks of this study.

4.2.1 RQ1: On the Ontology of the Mutable Work

The first research question, RQ1, was presented in Section 1.4.1 and investigates the ontology of the mutable work by asking:

Can high-dimensional latent spaces be structured to reflect perceptual, mathematical, and music-theoretical concepts, and what navigational paradigms allow these manifolds to function as intelligible, mutable musical works?

To address RQ1, we propose two sub-questions focused on the representation and navigation of latent spaces:

- **SQ1.1 (Structural Logic and Representation):** Are unsupervised latent spaces empirically aligned with formal music theory and perceptual sound characteristics to ensure that the mutable work maintains stylistic tonal coherence and timbre intelligibility?
- **SQ1.2 (Navigational Paradigm):** In what ways does the shift from discrete database querying to intent-aware topological navigation redefine the identity of a composition as a continuous manifold rather than a collection of fixed points?

4.2.2 RQ2: On Practical Applications and Contextual Navigation

The second research question, RQ2, was presented in Section 1.4.2. It investigates the performative dynamics and the redistribution of agency within the interface:

In what ways does the evolution of the epistemic tool (Magnusson, 2009) redistribute authorship, and how can contextual navigation be fostered through the machine’s material resistance?

To address RQ2, we propose two sub-questions centered around the epistemic role of the interface and the performative negotiation of agency:

- **SQ2.1 (The Epistemic Tool):** How does the implementation of real-time exploration interfaces transform the user’s interaction with musical logic from a task of execution to an epistemic process of discovery and navigation?
- **SQ2.2 (Contextual Navigation):** To what extent can the material resistance provided by the machine’s encoded contextual knowledge and high-dimensional latent space facilitate a state of contextual navigation, where authorship is redistributed between human and machine?

4.3 Operationalizing the Research Design

To address these research questions, this thesis employs a mixed-methods research design. This dual approach ensures that technical developments are subjected to rigorous experimental validation while simultaneously being refined through practice-based iteration. By adopting this framework, the study acknowledges that while formalized metrics can quantify many phenomena within a co-creative system, they do not capture the entire reality of the performative act. Therefore, the methodology is structured to bridge the gap between objective system benchmarks and subjective artistic experience.

4.3.1 The Experimental Track: Formalized Metrics and Validation

To satisfy the empirical requirements of RQ1, this research utilizes experimental methods to verify the ground truth of the developed latent spaces.

While the state-of-the-art (Chapters 2 and 3) acknowledges the existence of rhythmic latent spaces (e.g., GrooVAE), this research deliberately excludes the investigation of rhythmic and metric dimensions. This scoping is based on the ontological distinction between temporal sequencing and topological navigation. Rhythm, as a primarily horizontal and discrete temporal structure, relies on long-term dependencies and recurrent architectures that often resist the structural fluidity and continuous morphing explored in this study.

By focusing exclusively on tonal and timbre manifolds, the study prioritizes the investigation of vertical and qualitative dimensions, where the identity of the musical work is defined by its geometric position rather than its temporal sequence. This allows for a deeper exploration of

how formal musical logic (harmony) and sub-symbolic audio (timbre) can be unified into an intelligible, navigable interface, without the confounding variables of metric synchronization and rhythmic periodicity.

The experimental research track focuses on quantifying the relationship between symbolic (tonal) and sub-symbolic (timbre) representations and formal and perceptual musical logic, providing a baseline of technical stability and interpretability. This includes:

- **Mathematical Alignment (Representation):** Testing the correlation between unsupervised latent spaces and Fourier-based pitch representations. This quantifies the degree to which machine-learned abstractions mirror established music-theoretic and perceptual logics, addressing the structural requirements of the mutable work.
- **Dataset Empirical Verification (Intelligibility):** Using a prototypical corpus of J.S. Bach chorales, JSB corpus, and the produced SaxTimbre and Motiv datasets to quantitatively evaluate the system’s clustering performance. By measuring the distance between machine-generated points and human-defined taxonomies, we verify the system’s tonal and timbre intelligibility.

4.3.2 The Practice-Based Track: Iterative Development and Exploration

To satisfy the performative requirements of RQ2, the research employs a practice-based investigation, founded on practice-based research (Candy, 2006; Vear, 2022b). The developed systems (BroadcastJSB and Aethra) serve as primary research instruments in an iterative development cycle. While the experimental track provides the skeleton of the system, this track provides the flesh through:

- **Performative Prototyping:** Iterative refinement of navigation-based interaction through live performance. Each iteration acts as a probe into the latent territory, testing how theoretical constraints (such as latent regularizers) affect the tactile feel of the instrument.
- **Reflective Analysis:** Assessing the emergence of contextual navigation and the shifting role of the performer. This qualitative approach captures the phenomenological nuances of human-machine negotiation that metrics alone cannot reach.

4.3.3 The Iterative Research Cycle

The integration of these tracks is realized through a cyclic methodology where experimental results inform performative design, and performative insights identify new technical requirements for experimental validation. This approach ensures that the unified latent territory is not merely a mathematical abstraction but a functional site for contemporary musical practice.

1. **Phase One: Model Training and Latent Projection (Experimental):** Initial training of VAE (i.e., flat baseline recurrent VAE model and MusicVAE) architectures with curated

symbolic corpora and the projection of curated audio corpora into the latent spaces of pre-trained RAVE models. During this stage, the alignment with Fourier Phase Spaces and gestural taxonomies is measured to establish the representational explainability of the latent space.

2. **Phase Two: Performative Implementation (Practice-Based):** The validated models are integrated into the BroadcastJSB and Aethra systems. Here, the focus shifts to the interactional and conceptual levels of Explainable AI, testing the performer's ability to navigate the space in real-time.
3. **Phase Three: Performance and Reflection (Practice-Based):** Live realizations of the mutable works provide qualitative data on contextual navigation, agency, and the material resistance of the machine.
4. **Phase Four: Optimization and Dataset Refinement (Experimental):** Constraints and failures identified in performance lead to the refinement of the training datasets (SaxTimbre and Motiv) and the adjustment of latent regularizers to improve the navigability of the manifolds.

4.3.4 The Researcher-as-Navigator: Ethical and Methodological Positioning

By situating the researcher within this co-creative loop, the study adopts an auto-ethnographic lens common in practice-led research. As the primary developer and performer of the systems, my role is to act as the mediator between the machine's statistical logic and the aesthetic requirements of the musical work. This positioning allows for a thick description of the transition from a traditional composer-performer relationship to the navigator paradigm within a distributed agency framework.

Ultimately, this mixed-methods design provides a rigorous path from the abstract mathematical symmetries of tonal logic to the embodied, real-time improvisation of electroacoustic music. By validating the system's ground truth through experimental metrics and its artistic truth through practice, the thesis frames the mutable work as a comprehensive synthesis of human intuition and machine agency.

4.4 Implementation Roadmap: A Two-Phase Study Design

The mixed-methods approach is operationalized through two distinct phases of investigation, each targeting a specific dimension of the mutable work and addressing the refined sub-questions through a dedicated technical framework. In the following sections, we synthesize the experimental and practice-based track tasks for each phase of the investigation.

4.4.1 Phase One: Symbolic Tonal Representation (Part II)

Phase one addresses the structural and formal logic of the musical work. It investigates the transition from traditional symbolic notation to high-dimensional state spaces by focusing on the latent representation of tonal harmony.

- **Experimental Track:** The study employs variational autoencoders (VAEs), specifically MusicVAE (see Section 6.2.1), to model a corpus of J.S. Bach Chorales (JSB corpus, described in Section 5.3.0.1). The formalized metrics in this phase focus on the mathematical and music theory alignment between the machine-learned latent space and established music-theoretic models, such as the Tonnetz and Fourier Phase Spaces. We also focus on navigation strategies that allow us to navigate and generate music from these theoretically aligned spaces.
- **Practice-Based Track:** The findings are implemented in the BroadcastJSB system (see Chapter 8). This web-based interface enables real-time navigation of the learned tonal manifold, replacing prescriptive score-following with a model of topological discovery.

4.4.2 Phase Two: Expressive Timbre Navigation (Part III)

Phase two shifts the focus to the sub-symbolic and performative dimensions of the work. This phase investigates how the proposed contextual navigation model is fostered through the exploration of raw audio manifolds, where the interface acts as a sensory mediator between the performer's musical intent and the machine's contextually bounded possibilities.

- **Experimental Track:** Utilizing the RAVE (Realtime Audio Variational Autoencoder) architecture (see Section 9.1), this phase involves the development of the SaxTimbre and Motiv datasets, serving as the epistemic foundation for the navigable manifolds. Experimental setups focus on capturing extended saxophone techniques, providing a high-density timbre landscape for sub-symbolic modeling (SaxTimbre, see Section 9.2.2) and the encoding of musical gestures, defined here as intentional, musical phrase trajectories, into navigable latent coordinates (Motiv, see Section 11.3).
- **Practice-Based Track:** These technical developments culminate in the Aethra system (see Chapter 12), a co-creative environment for saxophonic performance. Through iterative live realizations, the researcher evaluates how motion-aware datasets transform the performer into a navigator of machine agency. In this context, the performative act is reframed as a continuous negotiation with the system's material resistance, effectively playing the latent space as a dynamic, sonorous object.

4.5 Summary of Research Methodology

This chapter outlined the mixed-methods research design of this thesis, aimed at bridging the gap between the mathematical potential of high-dimensional AI manifolds and their practical compositional and performative affordances. The methodology addresses three primary frictions: the opacity of AI's latent agency, the conflict between formal symbolic music theory and fluid sub-symbolic neural audio, and the complexity of mapping physical human motions to digital spaces. To resolve these issues, the study is guided by two core research questions focusing on the ontology and structure of the mutable work (RQ1) and the performative dynamics of contextual navigation (RQ2). The research relies on an iterative cycle combining an Experimental Track, which uses formalized metrics to objectively validate the tonal and timbre logic of latent spaces while deliberately excluding rhythm, and a Practice-Based Track, where the researcher auto-ethnographically acts as a navigator during the live prototyping of software systems.

The practical execution of this methodology is organized into a two-phase implementation roadmap. Phase one investigates the structural logic of symbolic tonal representation, utilizing VAEs (e.g., MusicVAE) trained on a corpus of J.S. Bach chorales to develop the BroadcastJSB interactive web application. Phase two shifts to the performative, sub-symbolic domain, exploring expressive timbre navigation through the RAVE architecture to build the Aethra co-creative performance system. To provide the necessary empirical foundation for these neural models, the study integrates the careful curation of specialized training data, specifically the SaxTimbre and Motiv datasets. Ultimately, by intertwining objective system benchmarks with subjective artistic exploration, this methodology provides a robust framework to transform the traditional performer into a navigator of an intelligent, co-creative ecosystem.

Part II

Latent Spaces of Tonal and Structural Dimensions

Introduction to Part II:

Latent Spaces of Tonal and Structural Dimensions

In the first part of this thesis, Part I, we established the concept of the musical work as a mutable interface. This second part focuses on how we can use tonal symbolic latent spaces as the underlying topological space for navigating within this mutable musical work.

As we have explored in Chapter 3, particularly in Section 3.2, latent spaces are high-dimensional, compressed mathematical representations of data learned directly by neural networks during the process of dimensionality reduction and reconstruction. Because of their powerful capacity for dimensionality reduction, they have shown great potential in generating meaningful and interpretable music latent spaces (Roberts et al., 2018; Guo et al., 2023). By mapping symbolic tonal music into these spaces, it becomes possible to explore variations and transformations that would be difficult to perceive in traditional score-based representations and descriptor spaces (Prang and Esling, 2020).

Thus, in Part II of this thesis, we focus on four interrelated objectives:

1. Representation of Tonal Structures:

Musical encoding has been proven essential to machine and deep-learning models' performance in symbolic music tasks, such as generation, transcription, and style recognition (Sarmiento et al., 2023). The encoding's properties determine the amount and quality of information that can be extracted from the data, directly influencing the accuracy and expressiveness of the generated output.

Therefore, our first objective is to determine whether the choice of musical encoding influences the perceptual alignment of a Variational Autoencoder's (VAE) latent space with cognitive tonal distance spaces. In Chapter 5, we will compare the latent spaces produced by different VAE corpus encodings (e.g., piano roll, MIDI, ABC, Tonnetz, DFT of pitch, and pitch class distributions) to see which best provides a tonal latent space aligned with cognitive key relations. This preliminary experiment utilizes a prototypical corpus of 371 chorales by J.S. Bach.

2. Mathematical and Conceptual Alignment:

Building upon the insight from Chapter 5 that representations based on the Discrete Fourier Transform (DFT) of pitch and pitch class distributions yield the most perceptually aligned latent spaces, we hypothesize that there is a fundamental correlation between VAEs’ latent spaces and theoretical DFT phase spaces.

Chapter 6 presents an experiment surrounding the same corpus of Bach’s chorales to mathematically quantify this geometric alignment. With this experiment, we attempt to determine if the VAE’s unsupervised learning process naturally captures formal music theory principles, such as common-tone functional harmony and parsimonious voice-leading, that DFT phase spaces are known to represent.

3. Sampling and Generative Exploration:

Having established that distances within the studied VAE latent space align with theoretical DFT phase spaces, our third objective investigates whether different distance-based strategies for navigating these spaces produce musically meaningful variations.

Chapter 7 experimentally compares three sampling strategies within the latent space trained in Chapter 6 to generate chord progressions mirroring Bach’s original sequences. Departing from an original Bach chorale phrase, we evaluate chord substitutions using: 1) discrete linear interpolation (LI), 2) k -nearest neighbors (k -NN), and 3) k -nearest neighbors combined with angular displacement (k -NN+AD).

4. Design and develop a topological instrument:

Guided by the conclusion in Chapter 7 that k -NN sampling strategies (particularly k -NN+AD) closely align with the tonal functions and voice-leading of the original chords, we are ready to operationalize the mutable interface.

To this end, Chapter 8 presents the design and development of BroadcastJSB, a web-based topological instrument. This system moves away from traditional “search-and-retrieve” generative tools toward a navigable territory where the performer can interact with the stylistic nature of J.S. Bach’s chorales in real-time. It implements a radio metaphor, serving as an intuitive cognitive bridge between the user and the underlying VAE latent space.

Through these explorations, we demonstrate how computational models can extend the creative agency of both composer and performer, allowing for interactive, flexible, and co-creative engagement with tonal structures. The chapters that follow integrate these theoretical, analytical, and practical perspectives, grounding them in concrete studies that form the backbone of the thesis’s investigation into mutable musical works.

Chapter 5

Latent Spaces of Tonal Music

This chapter looks at the challenge of effectively representing tonal music in compact, meaningful spaces that reflect both harmonic and phrase-level relationships. By employing variational autoencoders (VAEs), the musical sequences are encoded into latent spaces, where nearby harmonic and melodic structures can be found, fostering both analysis and generative exploration.

The goal of this chapter is twofold: the first goal is to show how these latent spaces expose structural aspects that are often tacit in Western tonal musical analysis, and the second is to illustrate their potential for creative practice. Through this lens, we hope to gather a new perspective on the material, enabling interactions that extend beyond traditional score-based constraints.

The remainder of this chapter is structured as follows. Section 5.1 introduces theoretical foundations connecting VAE latent spaces and tonal pitch spaces. Section 5.2 describes the methodology of our paper, presents each encoding’s implementation and characteristics, and our implementation of the VAE model. Section 5.3 presents our two-folded approach for evaluating the model performance, using typical reconstruction metrics, and the preliminary evaluation process to determine the effectiveness of latent spaces in representing cognitive distances between musical pitches. Section 5.4 presents and discusses the results of our two-folded evaluation approach, and, finally, Section 5.5 makes a last summary of the chapter.

5.1 Theoretical Foundations: VAEs and Cognitive Tonal Pitch Space

5.1.1 The Geometry of Tonal Pitch Cognition

The non-linearity and multidimensionality of humans’ pitch perception have motivated the search for a formal musical representation since the early days of musical computation. Early models by [Shepard \(1982\)](#) and [Krumhansl \(1990\)](#) demonstrated that musical distance does not simply rely on frequency intervals, but on cognitive proximity. This is particularly true in [Shepard’s](#) Tone Helix, where vertical proximity designates octave equivalence and the circle of fifths is represented by horizontal proximity.

Two historical lineages can be distinguished among these models. The first group focuses on the structural logic of musical composition and is based in music theory as well as the algebraic properties of the Tonnetz and harmonic voice-leading (e.g., [Weber, 1817](#); [Cohn, 1997, 1998](#); [Lewin, 2011](#); [Tymoczko, 2011](#); [Bernardes et al., 2016a](#); [Chuan and Herremans, 2018](#)). The second group, related to cognitive psychology, maps the mental representation of tonal hierarchies using empirical listener data, such as probe-tone experiments (e.g., [Shepard, 1982](#); [Longuet-Higgins, 1987](#); [Krumhansl, 1990](#)).

Both disciplines share a common goal: quantifying the closeness of pitch and key. For example, a meaningful space must mathematically represent why a listener perceives G major as closer to C major than to F# major.

5.1.2 VAEs as Manifold Learners

While these historical music theory and cognitive psychology-based models are derived either from mathematical formulations or cognitive models, respectively, VAEs are able to directly induce these tonal pitch spaces from raw data ([Kingma and Welling, 2014](#)).

According to Section 3.2 of the literature review on latent space models, we gathered that VAE models operate by mapping high-dimensional musical input x (e.g., MIDI-like piano rolls, with 128 possible activations per vector, or text-based ABC notation) into a low-dimensional latent bottleneck z through an encoder network $q_\phi(z|x)$. The VAE model must be able to strike a balance between producing a variety of unique, high-quality samples from a structured latent space and accurately reconstructing the data in order to be deemed well-trained.

The two main parts of the VAE architecture consist of an encoder, $q_\phi(z|x)$, which approximates the posterior distribution of the latent variables given the input data x , and a decoder, $p_\theta(x|z)$, which reconstructs the data from the latent representation such that:

$$\hat{x} = p_\theta(q_\phi(z|x)) \approx x \quad (5.1)$$

Unlike a standard Autoencoder, which may produce an overfitted latent space, the VAE architecture enforces a probabilistic prior, typically a standard Gaussian distribution, defined by $\mathcal{N}(0, I)$. For the probabilistic prior enforcement, the VAE architecture minimizes the Evidence Lower Bound (ELBO) ([Kingma and Welling, 2014](#)) objective function (Eq. 5.2). The first term of Equation 5.2 ensures reconstruction fidelity, ensuring the output matches the input, while the second term, the Kullback-Leibler Divergence (KL-Divergence), acts as a regularizer, by penalizing discrepancies between the learned distribution and a prior distribution (typically a standard Gaussian, $\mathcal{N}(0, I)$). This regularization is critical for musical applications because it forces the latent space to be dense and continuous.

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x) || p(z)) \quad (5.2)$$

By interpolating from this dense and continuous learned distribution, the model can generate new intermediate musical data instances that maintain the statistical characteristics of the training

corpus. In a well-trained musical VAE, empty regions between two encoded songs actually contain meaningful intermediate musical states, allowing for smooth interpolation between styles or melodies.

5.1.3 Bridging Latent Embeddings and Semantics

The VAE architecture constrains the sparse and discrete high-dimensional symbolic data into a low-dimensional manifold z . According to recent research (Roberts et al., 2018; Guo et al., 2023), the resulting manifold spontaneously reorganises itself in accordance with intrinsic musical hierarchies, such as pitch-class proximity or rhythmic density, when these models are well-trained on sufficiently large and representative corpora.

This convergence of the VAE's manifold with musical cognition allows the latent space to function as more than a simple storage mechanism. It becomes a generative grammar where spatial proximity serves as a proxy for perceptual similarity (Turker et al., 2022; Mezza et al., 2023). This enables three advanced types of semantic manipulations of the space: 1) attribute vector arithmetic, where it is possible to apply linear transformations to specific isolated concept vector sequences (e.g., major-to-minor mode change or diatonic-to-chromatic direction, while preserving the underlying rhythmic or structural identity); 2) semantic interpolation, through a cognitively plausible transition between two points in the latent space (i.e., two musical keys or textures); and 3) topological mapping and two- or three-dimensional visualization through dimensionality reduction techniques (i.e., UMAP, t-SNE, or PCA). When the resulting clusters mirror traditional musicological structures (i.e., the circle of fifths, or the Tonnetz), it provides empirical validation that the model has captured the fundamental cognitive distances described by Krumhansl (1990).

Despite retaining these capabilities, VAEs' latent spaces are often treated as black boxes. While Moss et al. (2023) and Buongiorno Nardelli et al. (2022) have successfully used data-driven models to track the evolution of tonality across history, Prang and Esling (2020) emphasize that this alignment with human cognition is highly sensitive to the initial data representation. The choice of encoding acts as a perceptual filter. If the encoding does not explicitly or implicitly capture interval relationships, the resulting latent space may fail to reflect the cognitive pitch distances central to human musical experience.

5.1.4 The Influence of Music Encoding on Latent Topology

The transition from theoretical cognitive models to a functional, generative system hinges on a fundamental design choice: the representation of data. As established, a critical factor in determining the semantic content of a learned latent space is the initial input representation, or encoding, of the musical data (Briot and Pachet, 2020). Because the VAE's objective is to compress information while maintaining reconstruction fidelity, it can only organize its manifold based on the features made explicit, or easily extractable, within the input format (Plitsis et al., 2020).

Certain musical characteristics, like absolute pitch, rhythmic duration, or relative interval relationships, are inherently highlighted by different encodings (Prang and Esling, 2020). The VAE

projects these features onto the latent space, directly echoing the bias of the encoding in the topography of the resulting manifold. When an encoding fails to capture the hierarchical nature of tonal relationships, the latent space will lack the cognitive alignment (Briot and Pachet, 2020) discussed in the previous section.

5.2 Experimental Design: Comparing Tonal Encodings

To determine the most effective encoding for capturing this tonal structure, our methodology involves a systematic comparison of VAEs trained on a unified corpus using various input formats (Carvalho and Bernardes, 2023). By holding the model architecture and the dataset (J.S. Bach’s 371 Four-Part Chorales, a canonic representation of the Western common practice period) constant, we can isolate the impact of the encoding itself. This approach enables us to quantify which representation best empowers the VAE to discover the latent geometric structures that mirror human musical cognition.

We compare typical VAE corpus encodings (piano roll, MIDI, ABC, and Tonnetz) against two proposed encodings based on spectral attributes of pitch, aligned with recent music theory practices using the Discrete Fourier Transform (DFT).

5.2.1 Methodology

The structural design of our comparative study is illustrated in Figure 5.1. To ensure a controlled environment for comparing symbolic music encodings, we developed a unified Python 3¹ framework designed to process disparate data formats under a single pipeline. This system leverages the music21² library to parse source material from diverse formats, including MusicXML, MIDI, and ABC, into a standardized internal representation.

To improve the model’s generalization and ensure a robust mapping of the tonal pitch space, we implemented comprehensive data augmentation. Each piece in the corpus is transposed across all 12 chromatic intervals for both major and minor modes. This process ensures that the VAE is exposed to the full spectrum of the circle of fifths, preventing the latent space from becoming biased toward specific, frequently occurring keys (e.g., C major). Detailed augmentation strategies tailored to the unique constraints of each encoding, such as maintaining the geometric properties of the Tonnetz or the periodic features of the DFT, are provided in Section 5.2.2.

The core of our methodology is the multi-encoding framework, which standardizes the training lifecycle across all six formats. By utilizing a shared pipeline for encoding extraction, prediction decoding, one-hot vectorization, and dataset storage, we isolate the encoding scheme as the sole independent variable.

¹<https://www.python.org/>, last accessed January 29, 2026.

²<https://www.music21.org/>, last accessed January 29, 2026.

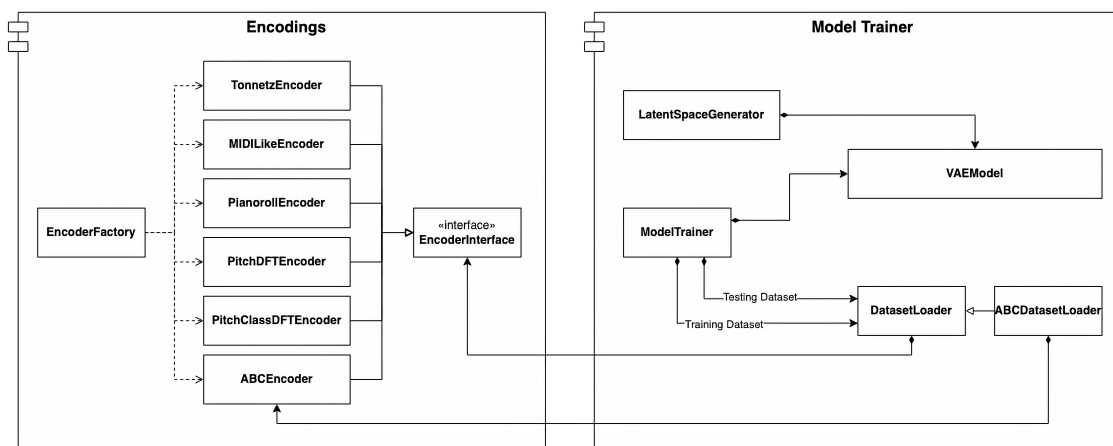


Figure 5.1: Schematic of the modular training pipeline. The training pipeline operates over an EncoderFactory pattern that handles representation-specific logic (e.g., one-hot encoding and decoding) while keeping a consistent ModelTrainer interface. This modularity enables the systematic evaluation of different musical encodings while keeping model hyperparameters and data-handling procedures constant.

To build our VAE implementation, including the definition of custom loss functions (incorporating both Reconstruction Loss and KL Divergence) and performance metrics, we use the TensorFlow (and Keras) framework.³ Section 5.2.3 presents the neural network architecture, hyperparameter selection, and optimization strategies specifics.

To make this research easier to reproduce and encourage more study of musical patterns, we have made our VAE architecture and multi-encoding framework publicly available. You can find the complete codebase,⁴ including preprocessing steps for six encoding methods and model training scripts. This repository acts as a guide for researchers, providing them with the tools they need to apply this method to other music datasets or creative tasks, including a Google Colab notebook.⁵

5.2.2 Symbolic Music Encodings

This section details the implementation and theoretical justification for each encoding scheme (piano roll, MIDI, ABC, Tonnetz, DFT of pitch class distributions, and DFT of the piano roll). To complement this analysis and facilitate further investigation, we provide an interactive platform.⁶ This tool allows users to explore the latent behavior and geometric properties of these encodings across a diverse range of symbolic music compositions, in uncompressed MusicXML format, providing a tangible bridge between the mathematical representations and their musical outcomes.

³<https://www.tensorflow.org>, last accessed January 29, 2026.

⁴<https://github.com/NadiaCarvalho/Latent-Tonal-Music>, last accessed January 29, 2026.

⁵https://github.com/NadiaCarvalho/Latent-Tonal-Music/blob/main/Latent_Tonal_Music.ipynb, last accessed January 29, 2026.

⁶Latent Tonal Music is available at <https://nadiacarvalho.github.io/Latent-Tonal-Music/>, last accessed January 29, 2026.

5.2.2.1 Piano roll (P)

The piano roll (P) is a high-resolution representation that consists of a binary matrix. In this matrix, each row corresponds to a MIDI pitch and each column to a discrete time step. For its elemental format, a value of 1 denotes note activation, while 0 indicates silence, as illustrated in Figure 5.2. Although this encoding technique is widely adopted for its simplicity and efficacy in representing both melodic and polyphonic structures, it suffers from a significant limitation, visible in the figure: the inability to distinguish between a single long sustained note and a sequence of repeated short notes.

To address this ambiguity, we implement an extended piano roll encoding, illustrated in Figure 5.3. This extension doubles the vertical dimension of the matrix, splitting the pitch space into two functional zones: 1) the first half of the encoding (128 dimensions for full MIDI) captures note-on events (attacks), marked 1 on the note's beginning timestep; 2) the second half represents active notes held over from previous timesteps (continuations). With the complete vector (i.e., concatenation of the two functional zones' vectors), the model can determine precisely the duration and articulation of each note. This ensures that the VAE reconstructs not only the correct pitch classes but also the rhythmic phrasing essential for the stylistic integrity of the J.S. Bach chorales.

We perform data augmentation by transposing the extended piano roll into all twelve chromatic keys. To that end, we implement a circular shift on the encoding vectors, in which the shift's magnitude is determined by the number of semitones involved in the transposition interval. This procedure is performed independently for the onset and continuation segments to preserve the structural integrity of the encoding. For the vector representing the notes' attacks, we rotate the entire vector (e.g., transposition C to D Major (+2 semitones) means a circular shift of +2 positions, with the activation bit moved from index 60 to 62, representing notes C4 and D4, respectively). Then, the vector representing the previous notes' continuations undergoes an identical rotation. By applying these shifts separately but synchronously, we ensure that the temporal link between a note's onset and its sustain is maintained, effectively remapping the entire musical sequence to a new tonal center without distorting the rhythmic or polyphonic relationships.

5.2.2.2 MIDI (M)

Our framework adopts an event-based MIDI protocol as the primary encoding method, following the methodology first proposed by Oore et al. (2020). Unlike the grid-based piano roll, this approach represents musical sequences as a chronological stream of discrete control messages. This MIDI-like encoding is particularly robust for handling complex polyphony and nuanced metrical variations, as it is not strictly bound to a fixed temporal grid.

The representation relies on a specialized vocabulary of four fundamental event types; Figure 5.4 illustrates the extraction process of this MIDI-like encoding from the raw source data. The vocabulary is partitioned into the following categories:

- **NOTE_ON**: Signals the start of a specific pitch (MIDI 0–127).

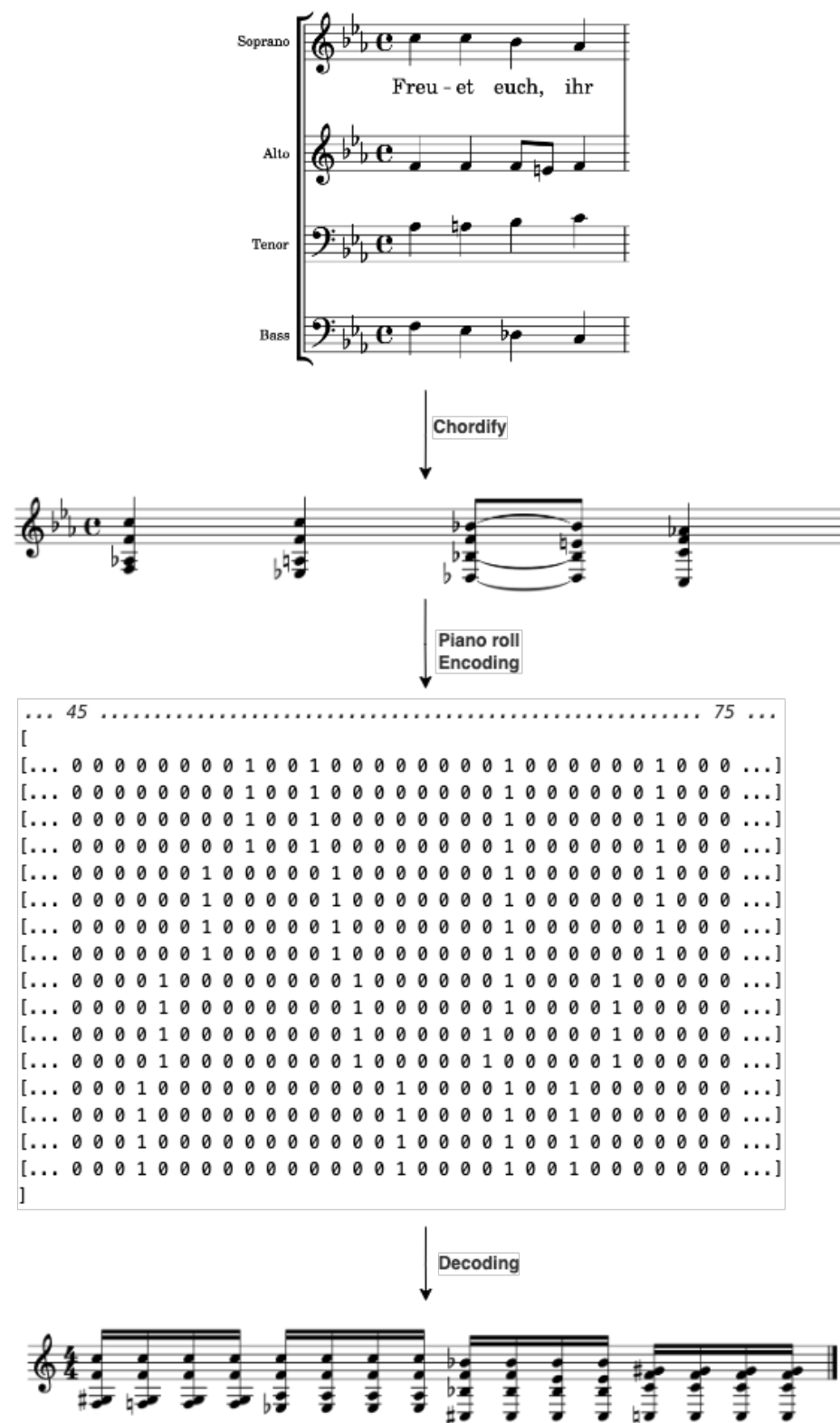


Figure 5.2: Process of encoding and decoding the first measure of *Freuet euch, ihr Christen alle* Bach (BWV 40/8) as a piano roll (Original encoding, 0-128).

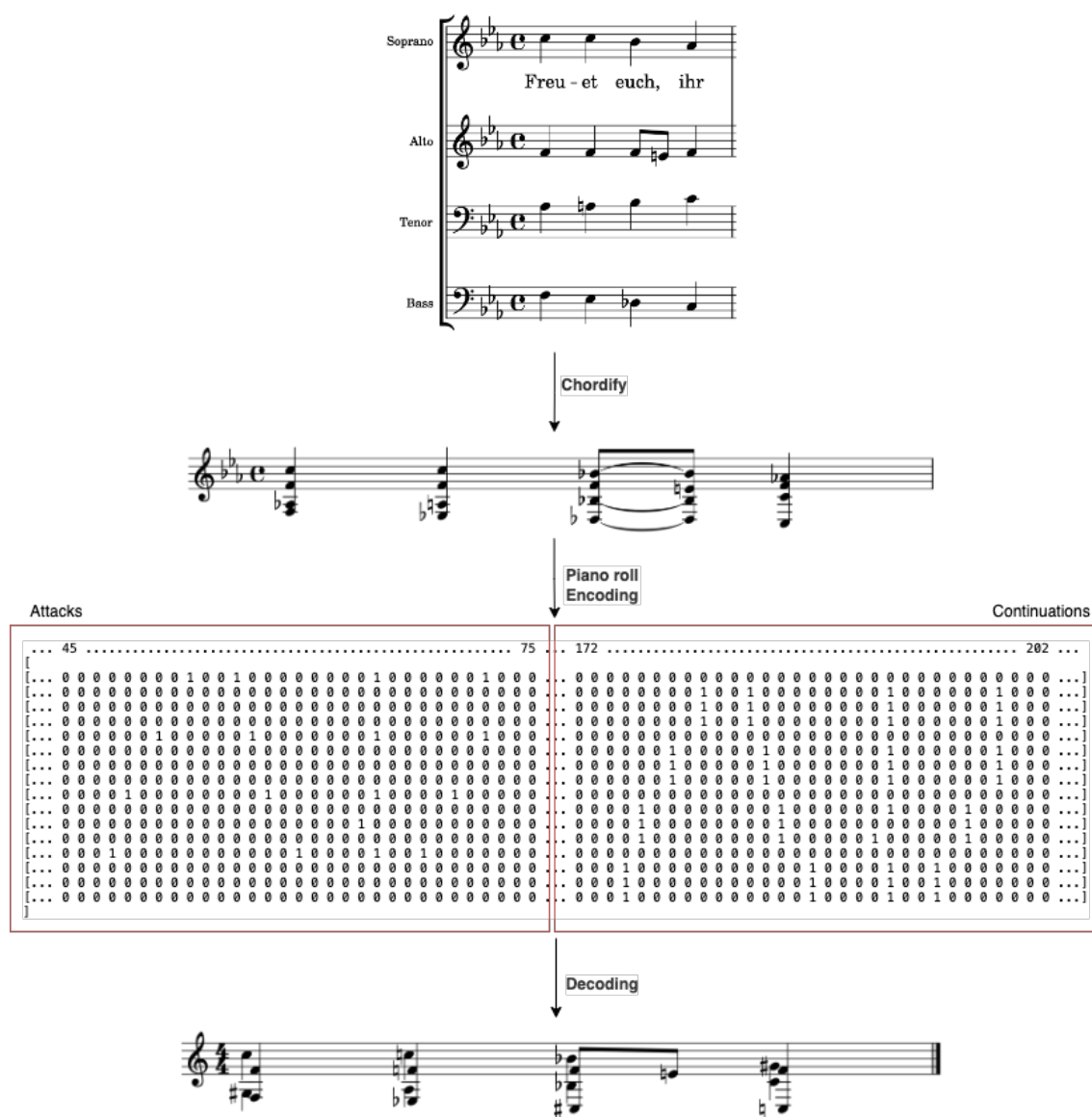


Figure 5.3: Process of encoding and decoding the first measure of *Freuet euch, ihr Christen alle* Bach (BWV 40/8) as a piano roll. (Using our continuation-based encoding, 0-128 for attack notes and 129-256 for continuation notes).

- **NOTE_OFF:** Signals the termination of a pitch.
- **SET_VELOCITY:** Defines the intensity or volume of the subsequent note.
- **TIME_SHIFT:** Advances the playhead by a discrete temporal increment, allowing for flexible rhythmic durations and micro-timings.

Because the MIDI-like encoding characterizes each timestep as a single categorical value within a vocabulary, the augmentation process is highly efficient. To transpose a sequence, we

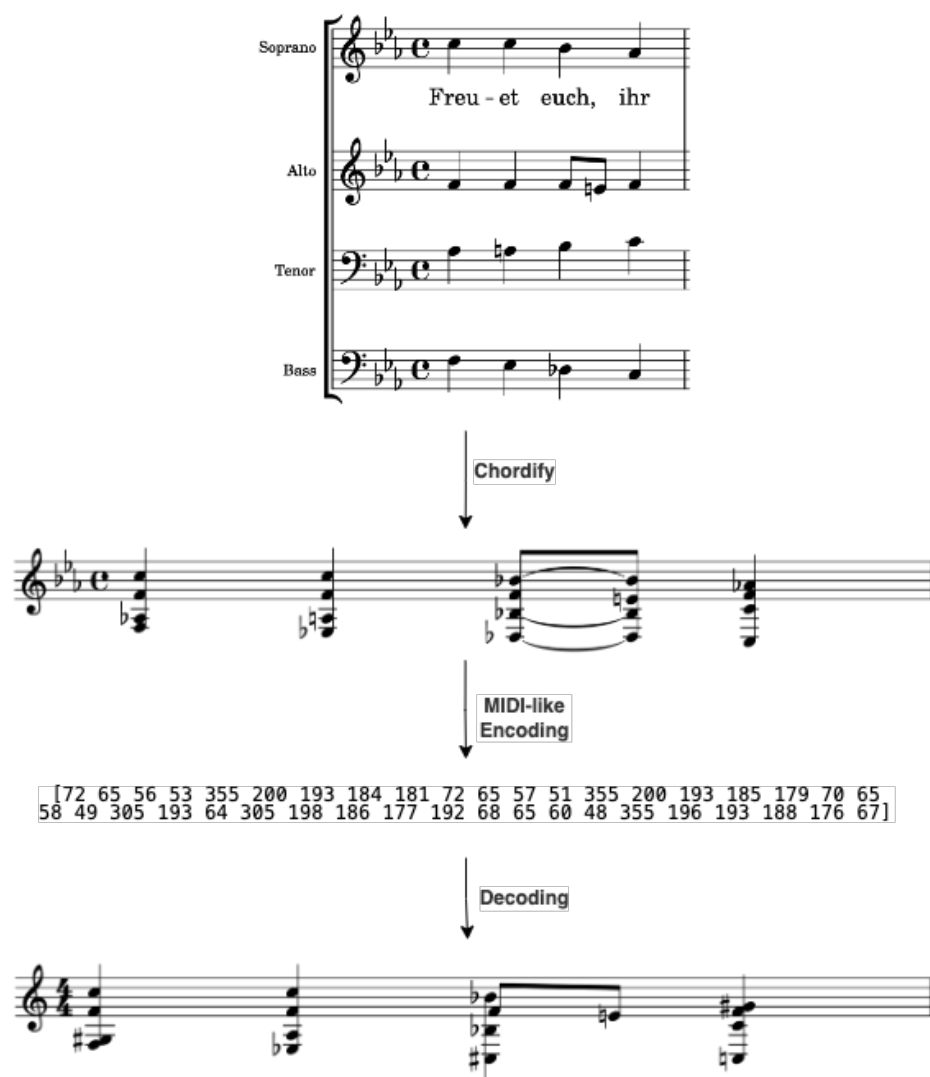


Figure 5.4: Process of encoding and decoding the first measure of *Freuet euch, ihr Christen alle* Bach (BWV 40/8) with MIDI-like encoding.

apply a numerical offset, corresponding to the number of semitones in the transposition interval, directly to the index values of all `NOTE_ON` and `NOTE_OFF` events.

When the data is converted into its one-hot encoded form for VAE processing, this transposition is mathematically equivalent to a horizontal circular shift of the vector. By rotating the activation bits by the same number of elements as the transposing interval, we maintain the structural relationships of the composition while shifting its tonal center.

5.2.2.3 ABC (A)

ABC notation is a shorthand textual system that uses alphanumeric characters and symbols to represent musical elements, including pitch, duration, and ornamentation. While traditionally valued for its human-readability and simplicity (Sturm et al., 2019), ABC notation offers a unique

linguistic perspective for machine learning models, treating music as a structured string of tokens rather than a numerical matrix.

Historically, ABC notation has been primarily applied to monophonic structures, such as folk melodies (Briot and Pachet, 2020). In this study, however, we adapt it as a textual representation for vertical harmonic aggregates (chords). To achieve this, we developed a custom parser based on the music21 library, drawing inspiration from the JavaScript Midi2ABC converter by Marmoo.⁷

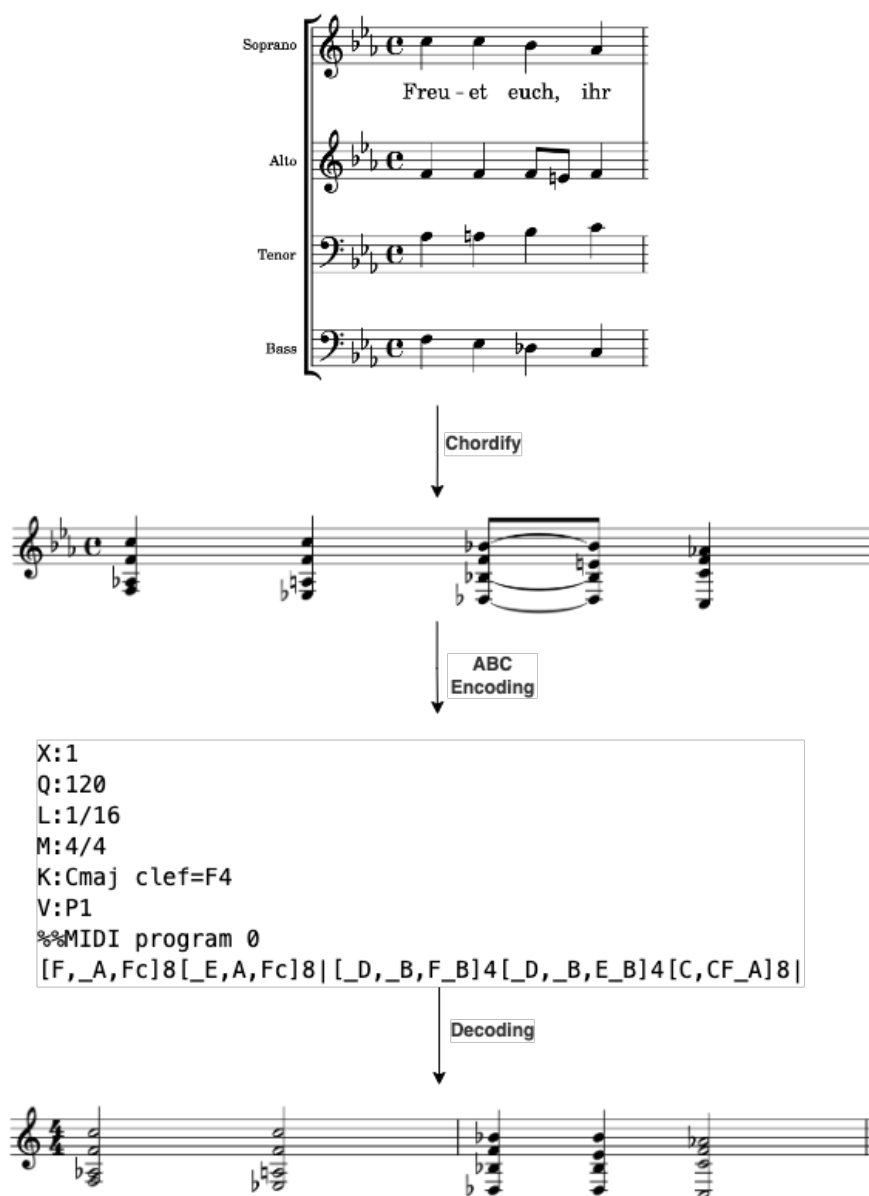


Figure 5.5: Process of encoding and decoding the first measure of *Freuet euch, ihr Christen alle* Bach (BWV 40/8) with the ABC encoding.

The extraction process, illustrated in Figure 5.5, involves several distinct stages:

⁷Implementation is available at <https://github.com/marmooo/midi2abc>, last accessed January 29, 2026. An online converter is available at <https://marmooo.github.io/midi2abc/>, last accessed January 29, 2026.

1. **Structural Segmentation:** The musical score is divided into instrument sections with corresponding headers.
2. **Harmonic Slicing:** Polyphonic textures are iterated through to convert chords and notes into ABC strings.
3. **Symbolic Translation:** Each event is translated into its ABC equivalent, strictly accounting for component pitches, note durations, tie marks, rests, and tuplets. This ensures the correct timing and chordal order is preserved in the character stream.

We acknowledge specific limitations in standard ABC notation regarding the preservation of passing notes within complex harmonic textures. To represent a polyphonic chorale in a single textual stream, a chordification process is often required. As shown in Figure 5.5, this can result in the loss of inner voice movement (e.g., an eighth-note passing tone in the tenor voice might be simplified into a block chord).

However, since our study prioritizes the cognitive distance of musical pitch within a key, the loss of minor rhythmic subdivisions is not considered a critical failure; it may, in fact, serve to aggregate non-chord tones into their primary vertical functional slices. To mitigate information loss, we implemented a strategy of splitting tied notes in chords where at least one voice moves. By treating these as distinct rhythmic events, we retain significantly more harmonic detail from the original Bach source while maintaining a valid ABC character sequence.

5.2.2.4 Tonnetz (T)

Building upon the theoretical geometry of the Tonnetz established in Section 3.1.1.1, this section details its adaptation into a functional machine-readable encoding. While it effectively maps theoretical tonal proximity, applying it as an encoding for deep learning requires significant structural modification. To capture the four-part polyphony of Bach’s chorales, we adopt the extended computational version introduced by Chuan and Herremans (2018), expanding it into a 24×12 matrix representation. Unlike the traditional pitch-class Tonnetz, this version maintains pitch register information (covering a range from C0 to C#8). The spatial arrangement is centered around a focal pitch, with register determined by proximity to the central column. This expansion is crucial for encoding the four-part polyphony of Bach’s chorales, as it enables the model to distinguish between different octaves of the same pitch class while still utilizing the horizontal circle-of-fifths relationship.

Following the logic of our extended piano roll, we further modify the Tonnetz representation by duplicating the matrix to capture both onsets (attacks) and sustains (continuations). Active notes are encoded as binary values (1 for activation, 0 for silence). As shown in Figure 5.6, a unique property of this encoding is that a single pitch may map to multiple nodes within the matrix due to its periodic nature. Consequently, for any active note, all corresponding coordinate positions in the Tonnetz must be activated simultaneously. This redundancy provides the VAE with multiple geometric viewpoints of the same tonal event.

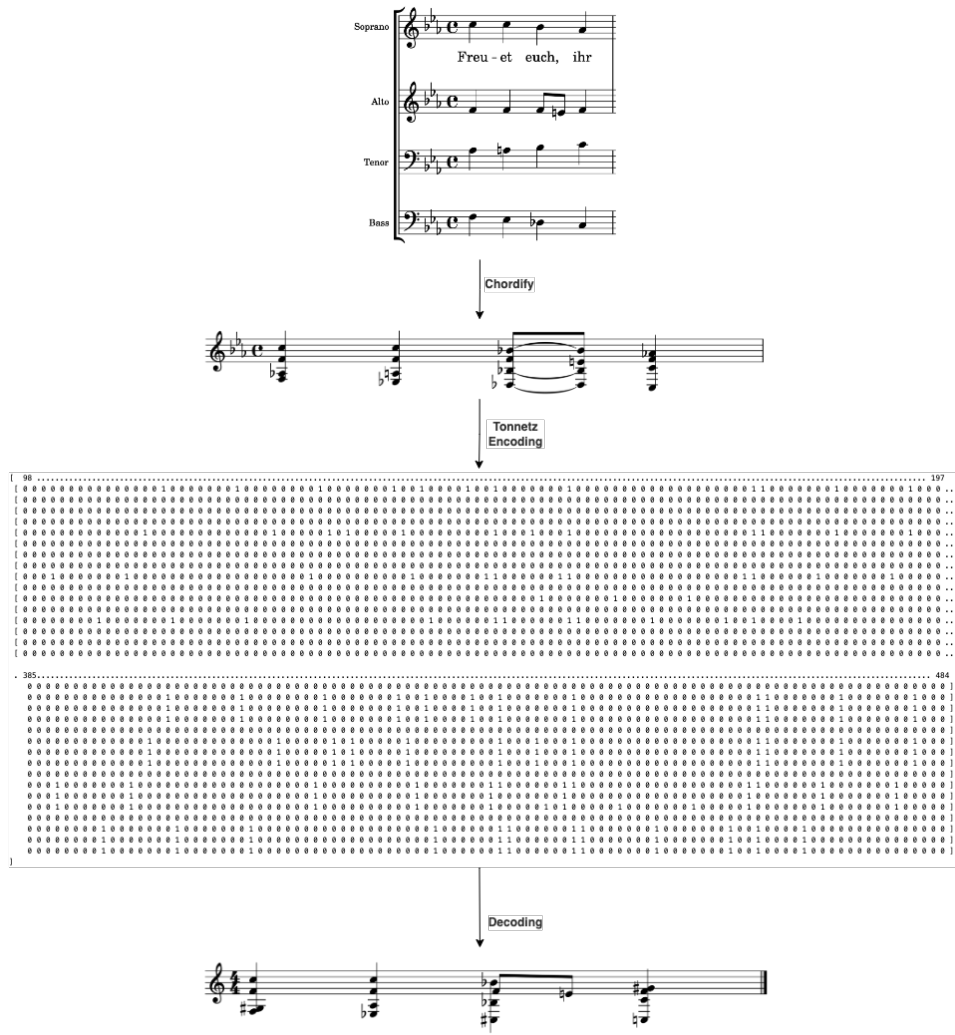


Figure 5.6: Process of encoding and decoding the first measure of *Freuet euch, ihr Christen alle* Bach (BWV 40/8) with the Tonnetz encoding. For legibility purposes, the continuations are below the attacks.

Due to the non-linear, grid-based mapping of the Tonnetz, performing data augmentation via simple vector rotation (as used in the Piano Roll) is not mathematically feasible. A shift in key requires a complete re-calculation of the active coordinates within the 2D lattice. Therefore, to perform our 12-key augmentation, each transposed version of the symbolic music is re-processed and encoded as a fresh Tonnetz matrix to ensure the geometric integrity of the harmonic relationships is preserved.

5.2.2.5 DFTs of Pitch Class (DFT-PC) and Pitch (DFT-P) Distributions

Drawing from research that utilizes Fourier space to describe musical objects and their intrinsic periodicities (Quinn, 2006; Yust, 2015a; Amiot, 2016; Bernardes et al., 2016a), we propose two novel encodings based on the Discrete Fourier Transform (DFT) of pitch distributions. These

encodings shift the representation from the time-pitch domain to a spectral domain, where musical features are expressed as complex coefficients.

A significant advantage of this approach is the separation of musical features: magnitudes ($\|z\|$) encode the interval content (the set-class of the pitches), while phases ($\angle z$) encode the tonal orientation or common-tone relationships between objects. We utilize the non-redundant output of the DFT ($m/2 + 1$ complex numbers) to represent these features efficiently.

The first proposed encoding, DFT-PC, reduces the distribution to 12 pitch classes ($m = 12$). As illustrated in Figure 5.7, we extract pitch class information into a 24-column array per timestep: the first 12 columns represent note attacks, and the subsequent 12 represent continuations.

The DFT is applied independently to the attack and continuation vectors. The resulting representation begins with a component reflecting the total number of activations, followed by the complex coefficients that describe the harmonic weights of the pitch set.

The second encoding, DFT-P, displayed in Figure 5.8, applies the DFT to the full pitch range ($m = 128$). We construct an extended piano roll of 256 elements (128 for attacks, 128 for continuations) and compute the DFT separately for each half. We retain only the non-symmetrical components (the first 65 coefficients for each part of the piano roll) and concatenate them into a singular vector. This preserves the register-specific information of the Bach chorales while exposing the intervallic periodicities across the entire pitch range.

To augment these encodings, we leverage the circular shift property of the DFT. In the spectral domain, transposing a musical sequence by an interval t does not alter the magnitudes, only shifting the phases of the complex coefficients.

To compute a transposition, we first determine the phase $a(z)$ of each component using the four-quadrant inverse tangent:

$$a(z) = \text{atan2}(\text{Im}(z)/\text{Re}(z)) \quad (5.3)$$

To transpose the encoding by t semitones, we rotate the angles of the original DFT components by t semitones. The new real and imaginary parts are calculated as:

$$\text{Re}(z+t) = \text{Re}(z) * \cos(t * a(z)) - \text{Im}(z) * \sin(t * a(z)) \quad (5.4)$$

$$\text{Im}(z+t) = \text{Re}(z) * \sin(t * a(z)) + \text{Im}(z) * \cos(t * a(z)) \quad (5.5)$$

This method allows for mathematically precise augmentation directly within the spectral representation, ensuring that the latent space learns the rotational symmetry of the circle of fifths.

5.2.3 VAE Model

While specialized architectures, such as MusicVAE (Roberts et al., 2018), utilize hierarchical conductor decoders to manage long-term dependencies, we opted for a flat baseline recurrent VAE model. This choice enables a more controlled comparative study, ensuring that the observed

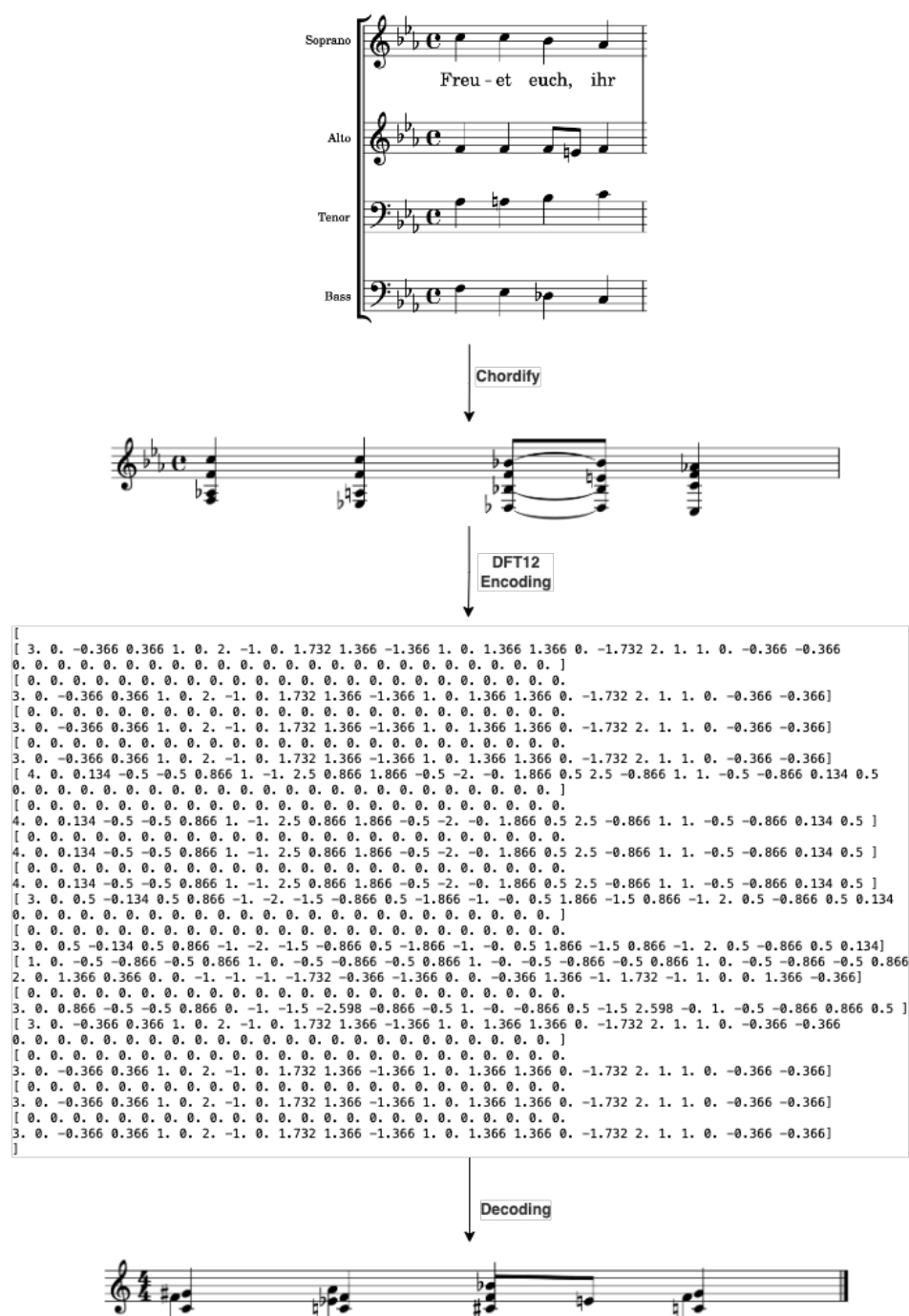


Figure 5.7: Process of encoding and decoding the first measure of *Freuet euch, ihr Christen alle* (BWV 40/8) with the encoding based on DFT from pitch class distribution (DFT-PC).

differences in latent topology are attributable to the encoding schemes themselves, rather than to architectural complexity.

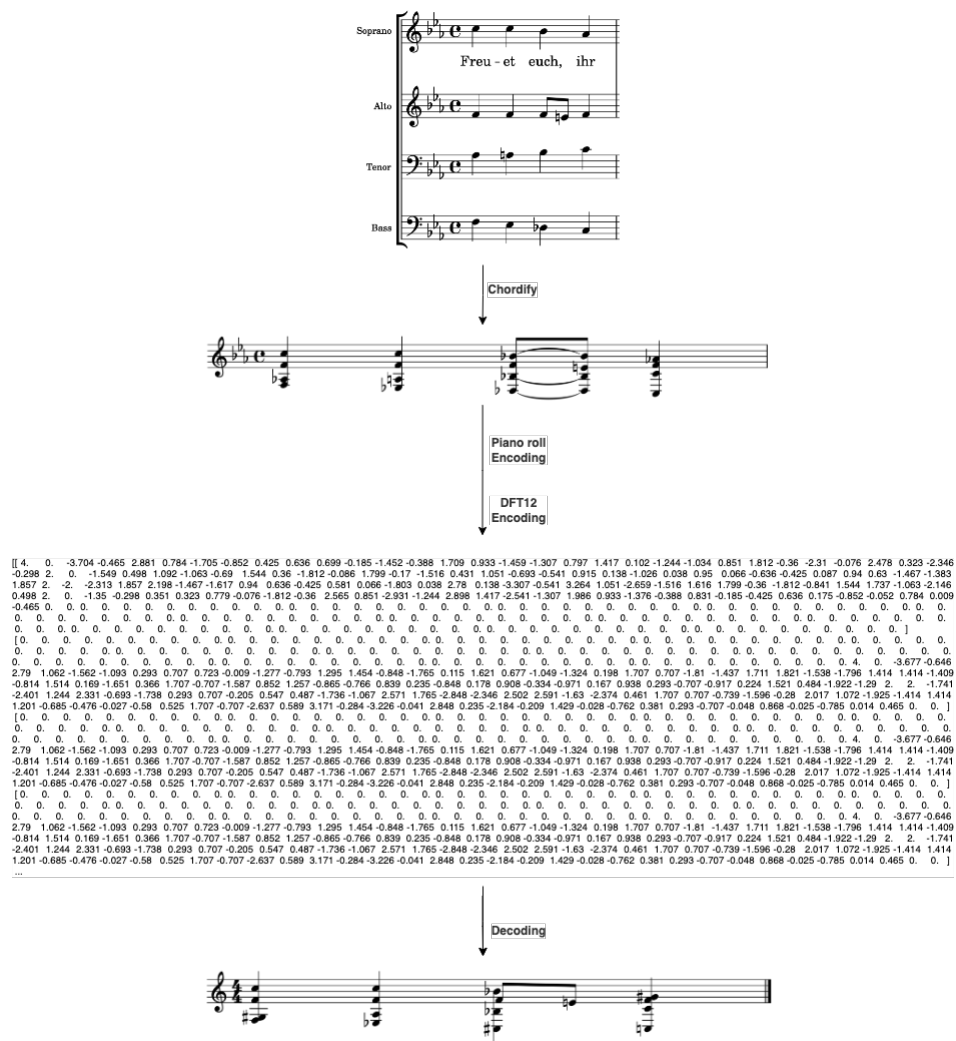


Figure 5.8: Process of encoding and decoding the first measure of *Freuet euch, ihr Christen alle* (BWV 40/8) with the encoding based on DFT from pitch (DFT-P). For legibility, we only show the output of the first four timesteps.

As illustrated in Figure 5.9, the implementation consists of two Long Short-Term Memory (LSTM) layers with 1024 units, serving as the encoder and decoder, respectively. We chose LSTMs specifically for their ability to capture long-range temporal dependencies and sequential patterns inherent in Bach’s polyphonic structures. Unlike standard recurrent units, LSTMs employ a gating mechanism, comprising input, forget, and output gates, that enables the model to maintain a persistent cell state. This is critical for music, as it enables the VAE to remember established harmonic contexts and rhythmic motifs across multiple measures.

The reconstruction loss term $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$ must be carefully matched to the numeric nature of the input representation:

- **Categorical Cross-Entropy (CCE):** Applied to the MIDI-like and ABC formats. These encodings treat music as a sequence of mutually exclusive tokens from a finite vocabulary.

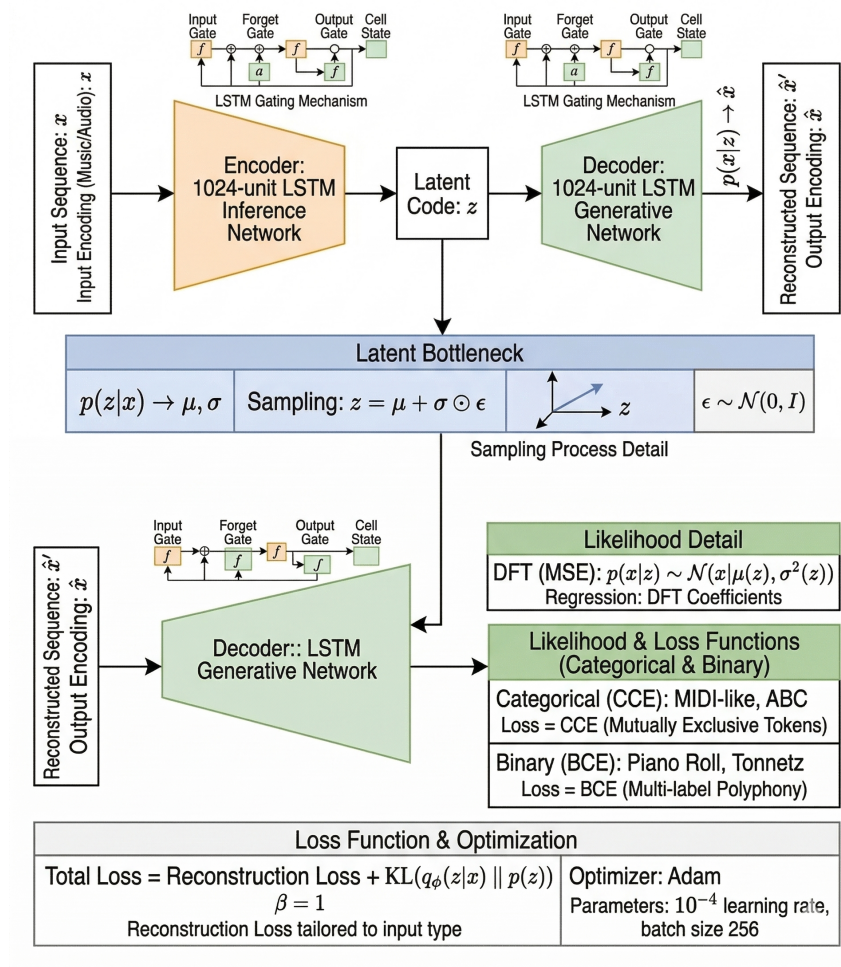


Figure 5.9: Schematic of the Implemented Flat Recurrent VAE. The architecture utilizes a symmetric design consisting of a 1024-unit LSTM encoder and decoder to capture long-range temporal dependencies in musical sequences. The latent bottleneck implements the reparameterization trick ($z = \mu + \sigma \odot \epsilon$) to enable gradient-based optimization. The reconstruction loss is dynamically adapted to the input encoding: Mean Squared Error (MSE) for continuous DFT coefficients, Categorical Cross-Entropy (CCE) for token-based MIDI/ABC formats, and Binary Cross-Entropy (BCE) for polyphonic Piano Roll and Tonnetz representations. The model is optimized using the Adam optimizer with a total loss comprising the reconstruction term and a standard KL-divergence ($\beta = 1$).

CCE measures the performance of a model whose output is a probability distribution over K categories, effectively penalizing the model for assigning low probability to the ground-truth token at each timestep (Goodfellow et al., 2016).

- **Binary Cross-Entropy (BCE):** Employed for the Piano Roll and Tonnetz encodings. Because these representations allow for multiple active notes per timestep (polyphony), they are modeled as a series of independent Bernoulli trials. BCE is the standard objective for multi-label classification tasks where each pitch’s activation is treated as an independent binary probability (Briot and Pachet, 2020).

- **Mean Squared Error (MSE):** Utilized for the DFT-based encodings. Since the Fourier coefficients consist of continuous-valued real and imaginary components (or magnitude and phase), we treat the reconstruction as a regression problem. MSE minimizes the squared Euclidean distance between the predicted and target spectral coefficients, which is essential for preserving the precise interval weights encoded in the magnitude values (Prang and Esling, 2020).

The total loss is the weighted sum of the reconstruction loss and the KL divergence, often controlled by a β parameter to prevent posterior collapse, where the model ignores the latent space in favor of the decoder’s own autoregressive power (Higgins et al., 2017). For this study, we maintain a standard $\beta = 1$ to observe the natural formation of the latent topology without external forcing.

As specified, the models were trained using the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 10^{-4} over a batch size of 256. This configuration ensures stable convergence across the various loss landscapes produced by our six encoding schemes.

5.3 Evaluation of Latent Space Alignment

The success of the latent space representation is quantified not only by the VAE’s ability to reconstruct the input data, but by the degree to which its topology reflects established cognitive musical principles. Consequently, our evaluation follows a twofold strategy designed to assess: 1) reconstructive integrity (primary evaluation), by measuring how effectively the VAE learns to compress and retrieve the source data; and 2) cognitive alignment to music theory principles (preliminary evaluation), by quantifying the degree to which the resulting latent space reflects the tonal relationships (e.g., the circle of fifths) established in music cognition.

5.3.0.1 Dataset and Preprocessing

To ensure structural consistency and a clear focus on common-practice tonality, we utilize a corpus of 371 J.S. Bach four-part harmonized chorales (JSB corpus). This dataset serves as a gold standard for tonal music research due to being a canonical representation of the Western common practice period.

The JSB corpus is obtained from the music21 library. It comprises 195 major-key (52.56%) and 176 minor-key (47.44%) pieces, as illustrated in Figure 5.10a. A significant portion of the total frequency is concentrated in just a few keys. The top three most frequent keys, G major (14.02%), A minor (12.13%), and G minor (11.86%), account for over one-third of all observed frequencies (38.01% combined), as seen in Figure 5.10b. Harmonic representation is dense, with an average of 84 chords per chorale. We employ a 60/40 train-test split, with the training set subject to 12-key chromatic augmentation to ensure tonal invariance in the learned representation.

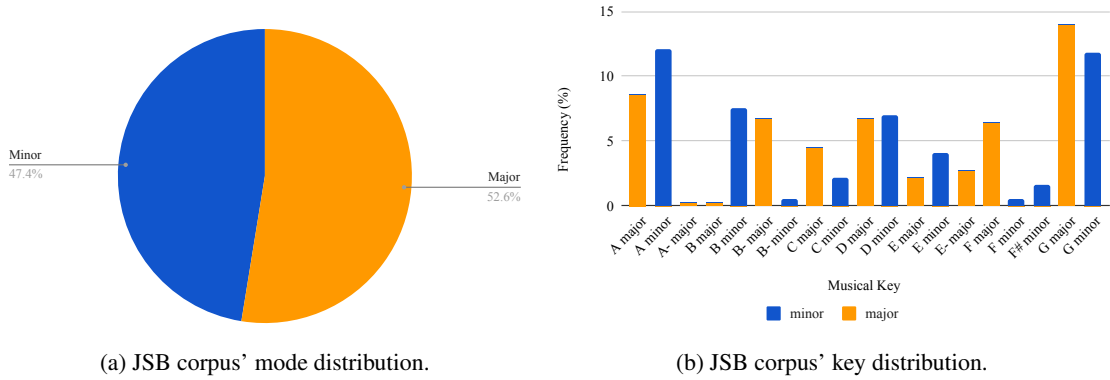


Figure 5.10: Distribution of musical (a) modes and (b) keys across the JSB chorale corpus. Blue color indicates minor keys and orange color indicates major keys.

5.3.0.2 Metrics of Model Performance

To evaluate the reconstructive fidelity of each encoding, we examined three core metrics: 1) accuracy, representing the percentage of reconstructed tokens that exactly match the ground truth (Briot and Pachet, 2020); 2) MSE, denoting the average squared difference between original and reconstructed sequences, crucial for continuous spectral data;⁸ and 3) KL-Divergence, which measures the discrepancy between the learned latent distribution and the Gaussian prior. Lower values indicate a reconstruction that adheres closely to the original data's probability distribution (Kingma and Welling, 2014).

5.3.0.3 Latent Space Topology and Cognitive Mapping

To make a preliminary evaluation of cognitive alignment of the latent space topology with music theory principles, we project 12 key transpositions of specific chorales, such as *Ich dank dir, lieber Herre* (BWV 347), into the latent space. We hypothesize that if the VAE has learned tonality, these transpositions should form distinct clusters whose spatial arrangement mirrors the circle of fifths.

We employ Principal Component Analysis (PCA) to project the 256-dimensional latent space into a 2D plane (Jolliffe, 2002). This metric enables maintaining the overall relationships and distances between tonal centers, which might get lost with non-linear methods (e.g., UMAP or t-SNE). To measure the musical intelligence of the clusters we create, we apply three different metrics:

⁸We used MSE as the loss function for DFT-based encodings to treat the reconstruction of Fourier coefficients as a regression problem. These coefficients include continuous real and imaginary parts (or magnitude and phase), so MSE helps us minimize the squared distance between the predicted and actual spectral coefficients. This step is important for keeping the correct intervallic weights in the magnitude values (Prang and Esling, 2020). Here, MSE serves as one of the main measures of performance for all encodings. It calculates the average squared difference between the original and reconstructed sequences, which is crucial for preserving the structure of continuous data in the latent space.

- **Davies-Bouldin Score:** This index checks how similar clusters are to their closest neighbors. For the JSB corpus, a lower score means the model has effectively grouped different keys into separate areas, which helps minimize confusion (Davies and Bouldin, 1979).
- **Dunn Index:** Complementary to Davies-Bouldin, this metric looks at the extremes by comparing the smallest inter-cluster distance to the largest intra-cluster distance. A higher Dunn index indicates that even similar tonal areas (e.g., relative majors and minors) are distinct enough for a performer to navigate them individually (Dunn†, 1974).
- **Kendall’s τ (τ):** Kendall’s τ is a circular non-parametric correlation coefficient (Fisher and Lee, 1982). In this case, we use this metric to measure the association between the ordinal sequence of keys in the latent space and the theoretical circle of fifths. As the circle of fifths is a closed loop, Kendall’s τ is able to determine whether the VAE can learn the cyclic nature of Western tonality (i.e., $|1|$ indicates a perfect structural mapping where the latent space is in perfect cognitive alignment).

5.4 Results and Discussion

Our evaluation highlights a key challenge: balancing computational efficiency with how well the models align with musical concepts. On one hand, our main assessment shows that the variational autoencoder (VAE) is great at retaining detailed data. On the other hand, our initial look at cognitive alignment reveals that models that are technically accurate don’t always display a deep understanding of music. In this section, we’ll explore how the six different encoding methods deal with the trade-off between the need for precise reconstruction and the need for a clear and meaningful latent space.

5.4.1 Model Performance and Computational Trade-offs

Table 5.1 presents quantitative results on the model’s reconstruction performance. The results indicated that the ABC encoding yielded the highest reconstruction accuracy and lowest KL-divergence, followed by MIDI-like encoding, suggesting that categorical one-hot formats are most efficient for data retention and lossless re-encoding by the VAE architecture. Conversely, piano roll and Tonnetz performed poorly in reconstruction, likely due to the sparsity of multi-hot vectors. For both representations, most of their representational space is filled with zeros, which makes the VAE allocate much of its bandwidth to recreating silence or inactive moments.

Interestingly, KL-divergence and MSE demonstrate consistent outcomes, reflecting a unified measure of performance, while accuracy does not reflect the same encoding order, with DFT encodings performing relatively well in accuracy but very poorly in the other two metrics. These results seem to suggest that while the model is successfully identifying pitches, with categorical accuracy metrics, it is not able to reconstruct the underlying continuous spectral coefficients with precision (i.e., DFT encodings’ very high KL-divergence values indicate the latent space is not

able to adjust into the smooth Gaussian distribution and is expanding to accomodate the high-dimensional complexity of the Fourier phases). We hypothesize that this could be a reflection of the reconstruction-regularization paradox (i.e., categorical encodings are easy for the model to memorize, leading to a well-compressed, low KL-divergence space). On the other hand, this dissonance in values for DFT spaces often signals a more informative and highly structured latent space (Ucar, 2019), with the model sacrificing global smoothness to preserve the intricate, music-theoretic relationships inherent in the DFT phase space. In terms of efficiency, the spectral (DFT) encodings offer a balanced profile, training significantly faster than the textual ABC format while maintaining competitive accuracy.

Table 5.1: Quantitative results on the VAE model’s reconstruction performance across the six presented encoding schemes: piano roll (P), MIDI-like (M), ABC (A), Tonnetz (T), DFT of pitch classes (DFT-PC), and pitch (DFT-P) distributions. ABC and MIDI-like encodings yielded the highest reconstruction accuracy and lowest KL-divergence.

Encoding	Acc. (%)	KL	MSE	Train Time/Epoch ^a	Embed. Time ^{a,b}
P	< 1	12.48	1.4×10^{-2}	~50 min	~2 min
M	77.4	0.80	8.0×10^{-4}	~150 min	~2 min
A	82.6	0.54	6.7×10^{-3}	~330 min	~3 min
T	< 1	35.29	1.9×10^{-2}	~55 min	~3 min
DFT-PC	77.3	136.02	6.5×10^{-1}	~70 min	~5 min
DFT-P	75.6	472.79	9.1×10^{-1}	~70 min	~20 min

^aAverage training time computed on a Macbook Pro M1 (8-core GPU).

^bTime required for feature extraction from the original symbolic source.

5.4.2 Latent Space Topology and Cognitive Mapping

Table 5.2 presents the average and standard deviation values for the cluster and key-distance metrics. This helps us preliminarily assess how well the musical encodings align with the cognitive high-level tonal abstraction of the circle of fifths.

Table 5.2: Average and standard deviation values for the cluster and key-distance metrics, assessing alignment with cognitive pitch spaces for the six presented encoding schemes: piano roll (P), MIDI-like (M), ABC (A), Tonnetz (T), DFT of pitch classes (DFT-PC), and pitch (DFT-P) distributions. DFT-P outperforms other encodings in all metrics.

Encoding	Davies-Bouldin	Dunn Index	Kendall’s τ (τ)
Piano roll	32.8 ± 17.8	0.0005 ± 0.0005	0.11 ± 0.08
MIDI-like	55.9 ± 35.3	0.0001 ± 0.0001	0.15 ± 0.14
ABC	66.9 ± 55.0	0.00003 ± 0.00003	0.11 ± 0.09
Tonnetz	33.0 ± 21.6	0.0006 ± 0.0007	0.11 ± 0.08
DFT-PC	37.3 ± 59.1	0.0006 ± 0.0006	0.11 ± 0.08
DFT-P	8.1 ± 5.5	0.0008 ± 0.0009	0.44 ± 0.32

The preliminary analysis revealed that the latent spaces derived from the DFT-P encoding exhibited superior alignment with cognitive tonal space, outperforming all other methods by a substantial margin. Its Kendall's τ value is significantly higher, indicating that the VAE is successfully organizing its latent space according to the circle of fifths.

Several notable patterns emerge from the DFT-P analysis. First, we identified a strong topological alignment in a subset of the data, where 16% of the chorales achieved a Kendall's τ absolute value greater than 0.9. This indicates a near-perfect circular arrangement in the latent space that mirrors the theoretical circle of fifths (see Figure 5.11).

Specifically, the DFT-based VAE was found to capture the hierarchical relationships of key components most effectively, providing a common-tone space where keys with high similarity. The chord slices sharing many common tones, such as parallel major/minor pairs or keys a fifth apart, are positioned closest together using this encoding. This result is significant, suggesting that spectral analysis of pitch content, rather than explicit note-event modeling, yields a latent structure that most accurately mirrors the human perception of tonality.

Furthermore, the results reveal a distinct mode bias, where major-key chorales were overwhelmingly more successful in generating highly organized latent spaces; 87% of the pieces reached the $|\tau| > 0.9$ threshold, and all of these were in a major key. This suggests that the model finds the internal logic of the major scale more mathematically consistent. In contrast, the minor mode's inherent complexity, often involving variable sixth and seventh degrees (natural, harmonic, and melodic minor), may introduce a degree of harmonic noise that makes the global organization of minor keys in the latent space more challenging to discern.

Finally, we observed a correlation between composition length and structural clarity. Counter-intuitively, longer chorales (i.e., exceeding 114 slices) demonstrated superior alignment compared to shorter pieces. We speculate that these longer compositions provide a greater volume of tonal evidence, which effectively assists the VAE in resolving key center ambiguities. While shorter fragments might be tonally ambiguous, the extended duration allows the model to capture a more complete spectral fingerprint of the key, helping it navigate modulations and establish a stable position within the latent space.

Ultimately, while textual encodings like ABC are superior for verbatim reconstruction, spectral encodings (DFT) are the most effective for inducing a musicologically intelligent latent space that mirrors human perception.

5.5 Summary: Musicological and Creative Implications

The finding that a latent space derived from the DFT-P encoding aligns most closely with cognitive models offers profound implications for both analytical and generative musicology. By demonstrating that a VAE can discover the circle of fifths through spectral representation, this work bridges the gap between deep learning architectures and centuries of music theory.

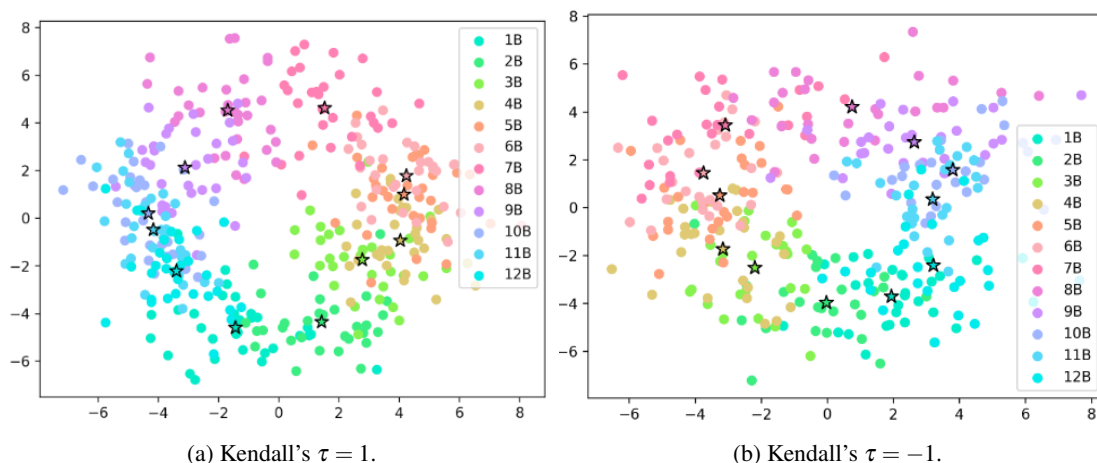


Figure 5.11: Latent space projections for two chorales using the DFT-P encoding across twelve key augmentations. (a) Displays a perfect positive circular correlation (Kendall's $\tau = 1$). (b) Shows a perfect negative correlation ($\tau = -1$). In both cases, the latent space preserves the circle-of-fifths topology, differing only in the clockwise versus counterclockwise orientation of the key clusters. Cluster coloring and numbering follow the Camelot Wheel convention (keys B through E represented by indices 1–12, with 'B' denoting the major mode). Colored stars indicate the calculated centroids for each key cluster.

5.5.1 Analytical Contributions

Analytically, the latent space provides a low-dimensional map where the trajectory of a musical work can be visualized in relation to established key areas. This enables the quantification of harmonic tension, modulation distance, and formal structure, features that are often subjectively interpreted in conventional Schenkerian or Roman numeral analysis.

The emergence of a circular, continuous topology suggests that we can now use VAEs as computational mirrors of human tonal perception. Researchers can move beyond simple classification (e.g., “Is this in C major?”) to analyze the nuances of a piece’s position within the tonal manifold, tracking how a composer navigates the fuzzy boundaries between related keys.

5.5.2 Generative and Creative Potential

Generatively, this semantically grounded space facilitates exploration through interpolation and manipulation along meaningful linear directions. Traversing a path between the latent vector of a piece in C major and one in G major provides a computationally informed method for creating smooth, tonally coherent modulations.

This capability extends creative practice by offering composers an intuitive, geometric interface for interacting with musical material. Rather than manipulating discrete notes, composers can sculpt the harmonic and melodic flow of a sequence as if it were a physical manifold. This shifts the role of the AI from a mere imitator of style to a cognitive partner that understands the underlying tensions of the Western tonal system.

5.5.3 Summary

While our study focused on the prototypical tonality of J.S. Bach, the success of the DFT-P encoding suggests that this framework could be applied to more complex or less studied harmonic systems, such as modal jazz, non-Western microtonal scales, or contemporary atonal structures. If the encoding can expose the relevant periodicities of a scale, the VAE can likely induce its corresponding cognitive space.

This chapter's findings highlight the importance of correctly choosing a VAE's latent space encoding and how it can reflect our theoretical views on what artificial intelligence can understand about music. In our earlier findings about how generative models align with human thinking (Carvalho and Bernardes, 2023), we discovered that by focusing on spectral information, we might be able to develop systems that don't just mimic a musical score but also honor the deeper emotional aspects that shape the human experience of music.

Ultimately, the emergence of a common-tone space through the encoding based on the DFT-P suggests that the mathematical periodicities of the Fourier transform provide a robust bridge between signal processing and music cognition. This alignment opens new avenues for AI-assisted musicology, in which we can exploit latent spaces as a navigable, geometrically intuitive map of tonal relationships that can inform both rigorous analysis and creative practice.

By revealing the rotational symmetries and topological continuities inherent in pitch-class distributions, we have established a framework where musical meaning is encoded as spatial distance. This provides the necessary foundation for the following chapter, which moves beyond representational accuracy to examine the formal mathematical correspondences between these learned models and Fourier-based phase spaces. We will explore how the geometry of these latent spaces reveals deep-seated patterns of harmonic motion and symmetry, further reinforcing the concept of the musical work as a flexible, interactive system.

Chapter 6

From Fourier Phase Spaces to Latent Representations: Mathematical Structures in VAEs

The previous chapter preliminarily established that the Discrete Fourier Transform (DFT) of pitch content yields a VAE latent space that aligns strongly with cognitive models of tonal proximity (Carvalho and Bernardes, 2023). This discovery suggests that we explore further into the math behind this alignment. Specifically, we should look into how the way a variational autoencoder (VAE) organizes its learned space connects to the structures we know from Fourier analysis in theoretical phase spaces. While traditional neural networks have shown high correlations between internal weights and DFT phases, our work extends this investigation to the interpretable latent spaces of generative models.

In this chapter, we propose relying on a more advanced VAE model, the MusicVAE architecture, to demonstrate how these latent spaces expose formal theoretical structural aspects of musical harmony. To that end, we use the representative Western tonal music corpus of J.S. Bach’s four-part harmonized chorales (JSB corpus). We train the MusicVAE model on this musical corpus, represented as piano roll encodings.

The chapter’s outline starts by defining the adopted theoretical DFT phase spaces (Section 6.1). Then, Section 6.2 delineates the methodological approach for our experiment, which intends to prove whether there is a correlation between pitch and pitch distribution DFT phase spaces and the trained latent space. Subsequently, in Section 6.3, we offer the results of the experiment, discussing implications in the latent space’s ability to capture tonal function and voice-leading aspects, and stylistic signatures of the JSB corpus. We discuss a representative case study around the *BWV 307* chorale, which showcases the findings of our experiment. At last, Section 6.4 concludes the chapter, offering a summary of our experiment and its principal conclusions, pointing towards a structurally hybrid AI paradigm that allows for clearly explaining musical concepts.

6.1 Defining Theoretical DFT Phase Spaces

In Section 5.2.2.5, we presented symbolic Fourier phase spaces as encodings to train VAE models. We showed how they can be obtained from either pitch or pitch class distributions, with the last one being the more conventional approach in the literature. Typically, each of these DFT phase spaces intends to highlight different aspects of tonal function and voice leading. In this sense, to establish a basis for comparison, we construct our experiment using both phase spaces, which we will define in the following sections.

6.1.1 DFT Phase Space of Pitch Classes

In music theory and music information research, Fourier phase spaces are commonly derived by applying the DFT on a pitch class set represented as a binary distribution over 12 elements. In this representation (DFT $\{0, 1\}^{12}$), unit activation denotes the presence of a pitch class (i.e., a C Major chord [60, 64, 67, 72] would be represented as the vector [1, 0, 0, 0, 1, 0, 0, 1]). This chroma representation reduces chord inversions and octave displacements to an agnostic space, focusing on functional harmonic clusters.

When inputting this vector into the DFT, we get 12 complex-valued DFT coefficients (c_k). As we saw in Section 5.2.2.5, the DFT components are symmetric, so we only need coefficients one to six, from which we can encode magnitude and phase information (Quinn, 2006). The magnitude of the coefficients is typically related to harmonic qualities. On the other hand, the phase reflects the pitch organization within regional sets, resulting from common-tone pitch proximity arrangements (Yust, 2015b). Coefficients $k = 3$ and $k = 5$ are important in the literature. The phase of these coefficients is usually associated with spaces shaped like a torus (e.g., the Tonnetz) (Amiot, 2013; Yust, 2018), capturing triadic voice-leading and tonal region relationships (e.g., the circle of fifths and common-tone functional harmony).

Because of this geometrical capability, the space constructed from the phase information of these coefficients is very relevant in the literature, as it is possible to effectively project a wide range of harmonic objects (e.g., pitch, chords, and regions) within established theoretical hierarchies. Within this continuous geometrical space, functional harmonic clusters become highly visible, as chords sharing similar tonal functions are positioned in proximity to one another. Adopting L_1 pitch class distributions normalization, we can compare and analyze different hierarchies (Bernardes et al., 2016a), aligning with many music-theoretical and cognitive spaces of interest. Practical applications of these Fourier phase spaces have been diverse across music theory and creative practice:

- **Analytical Study:** Yust (2015b) utilizes the third and fifth coefficients to construct a two-dimensional phase space, studying the Trio and Adagio from Schubert’s String Quintet (D. 956), and refining neo-Riemannian concepts to analyze chromatic harmony and triadic voice-leading.

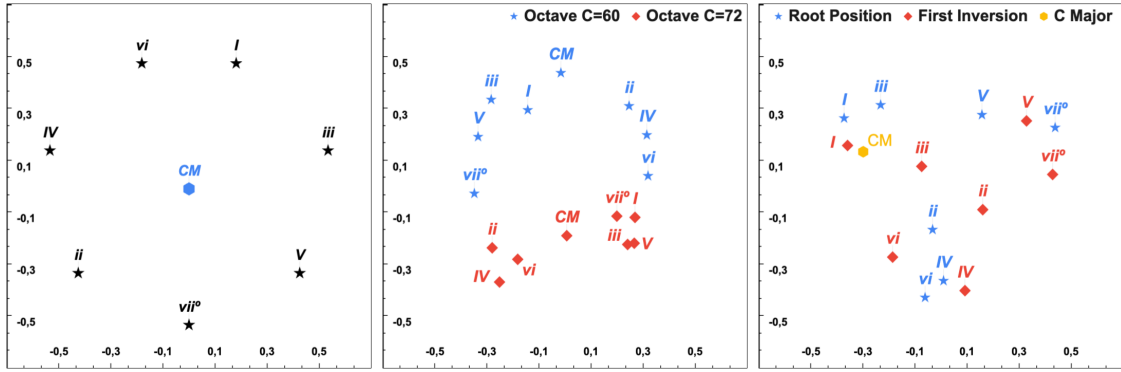


Figure 6.2: Two-dimensional LLE visualizations of DFT phase spaces for diatonic chords in C major. The figure depicts DFT $\{0, 1\}^{12}$ (left), DFT $\{0, 1\}^{128}$ (center and right), showcasing sensitivity to octave displacement (center figure, where identical chords are displayed across different octaves) and chord inversions (right figure, where chords are shown in root position and first inversion).

As shown in Figure 6.1, both the pitch class and pitch distribution spaces exhibit complete alignment when representing major and minor regions. However, as illustrated in Figure 6.2, while functional harmonic clusters (e.g., C major diatonic chords) are clearly observed in the DFT $\{0, 1\}^{12}$ space, they are less represented in the pitch distribution space. For example, in the DFT $\{0, 1\}^{128}$ space, C major and A minor chords do not align with any common tone as they do in the pitch class space. Despite this, the pitch distribution space more effectively expresses parsimonious voice-leading. The relations between articulated chords in this space denote proximity based on tone retention.

6.2 Methodology: Correlation between Latent and Phase Spaces

This chapter looks into how these theoretical phase spaces mathematically correlate with the latent representations learned by VAE models. We hypothesize that VAEs, particularly MusicVAE, despite being trained on raw musical data, naturally correlate with DFT phase spaces. Such a correlation suggests that these generative models effectively capture common-tone principles, yielding organized spaces where functional harmony and voice-leading can be interpreted.

While the previous chapter demonstrated the topological advantages of using an encoding based on the DFT of pitch distributions, the current experimental framework utilizes a binary 128-element piano roll encoding for the VAE training. The objective is to determine if a model trained on this standard, non-theoretical representation can internally derive a latent structure that mirrors formal Fourier-based phase spaces. To make it easier for others to reproduce our results and continue this investigation, we share the evaluation process code developed, including the calculation of distance matrices and correlation analysis.¹

¹<https://github.com/NadiaCarvalho/MCM-Latent-Spaces>, last accessed January 29, 2026.

6.2.1 Model Architecture and Training

In this experiment, we used the MusicVAE model, designed to understand complex musical patterns thanks to its unique hierarchical structure. Unlike the model used in Chapter 5, which treated music as a simple flat representation, MusicVAE breaks down musical data into smaller sub-embeddings, using a conductor-decoder architecture (Roberts et al., 2018). This layered approach helps the model capture the various levels of musical structure. Therefore, the model performs better at representing long-term structural relationships and common-tone transitions, which simpler models might overlook.

MusicVAE’s architecture employs a recurrent encoder, featuring a two-layer bidirectional LSTM with 1024 units. The decoder is structured hierarchically, employing a two-layer unidirectional LSTM with a hidden size of 1024 for both the conductor and decoder components. A latent size of 256 was chosen to significantly diminish input dimensionality while retaining sufficient information for effective reconstruction (Prang and Esling, 2020). Training was conducted using the Adam optimizer with an initial learning rate of 10^{-4} and a batch size of 16.

6.2.2 Dataset and Augmentation

Similar to the experiment detailed in Chapter 5, we employ the benchmark prototypical tonal corpus of J.S. Bach’s four-part harmonized chorales (for more information on the JSB corpus, see Section 5.3.0.1). In the previous chapter, we found that ABC encodings are better for reconstruction, and DFT from pitch encodings are better for cognitive alignment. However, for this experiment, we are using the traditional 128-element binary piano roll (without the continuation extension proposed in Section 5.2.2.1). We chose this representation to inspect whether MusicVAE can independently uncover the mathematical patterns of Fourier phase spaces. This encoding helps us maintain important details like chord inversions and voice leading, which can be lost in symbolic formats like ABC. By doing this, we can better connect with the theoretical DFT models.

In this experiment, we decided to create a custom training method instead of using pre-trained MusicVAE models. This allows us to keep a close connection between the model’s input and the theoretical DFT distributions. Most standard MusicVAE models used complex MIDI formats and were trained on a mix of different music styles, like the Lakh MIDI dataset. By focusing our training only on the JSB corpus and using simple binary piano-roll encodings, we can make sure that the model reflects clear tonal principles without the extra complexity that comes from different music genres or variations in rhythm.

Once more, to ensure MusicVAE develops a generalized understanding of tonal relationships rather than memorizing specific key frequencies, the following preparation steps were taken. The corpus was partitioned into a 60% training set and a 40% test set. To avoid the model favoring predominant keys like G Major or A minor, we made sure to transpositionally augment the dataset across all twelve keys. By doing this, we ensured that every key got fair representation and kept training MusicVAE balanced. Similar to the corpus processing described in Section 5.3.0.1, we use

“salami slice” segmentation, expecting to minimize the impact of rhythmic information, allowing the VAE to focus primarily on the structural organization of pitch.

6.2.3 Evaluation Methodology

To construct the three distinct geometric representations (MusicVAE latent space, DFT $\{0, 1\}^{12}$, and DFT $\{0, 1\}^{128}$ spaces), basis for our comparison, we follow an equivalent procedure. Each slice of the musical data is transformed into its respective representation by passing through either the MusicVAE encoder to obtain a latent vector (z) or the appropriate DFT process to generate a theoretical phase vector (c). This methodology allows for a direct comparison between the model’s learned geometry and theoretical constructs.

We establish the alignment between these spaces by calculating pairwise distance matrices for every slice in the test fold. For the latent space, we employ both Euclidean and cosine distance metrics. Cosine distance is utilized only for the DFT spaces to specifically capture phase differentiations. The resulting distance matrices are then compared pairwise using Pearson (r) and Spearman (ρ) correlation coefficients.

The use of the Spearman rank correlation is particularly significant as it assesses the monotonic relationship between the spaces. Rather than requiring a strict linear correspondence, this non-parametric measure determines if the relative ordering of distances between musical segments is preserved. When this happens, it indicates that MusicVAE has effectively captured the hierarchical structural relationships inherent in the theoretical DFT space. Correlation values approaching one signify a strong preservation of these theoretical harmonic relationships within the learned model.

Beyond global regional alignment, this methodology also assesses harmonic hierarchy at the functional level. While the correlation of distance matrices provides a quantitative measure of regional proximity, the specific case study of *BWV 307* allows us to evaluate the micro-topology of the latent space. In this case study, we examine whether MusicVAE is able to cluster chords based on their harmonic roles (e.g., Tonic (T), Subdominant (SD), and Dominant (D) functions), and how these ‘functional clusters’ compare to the theoretical boundaries defined by the DFT phase spaces.

6.3 Results and Discussion

The experimental results demonstrate a reasonable linear and rank association between the MusicVAE latent space and the theoretical DFT phase spaces. Two primary observations can be drawn from the data presented in Table 6.1, which displays the Pearson r and Spearman ρ correlation values for the corpus’s latent spaces and DFT phase distances.

First, by performing consistently better using the cosine distance for latent space correlation, our results seem to indicate that MusicVAE’s tonal logic is primarily encoded in the angular orientation (phase) of its vectors rather than their magnitude (e.g., vector length or L_2 norms). However, based on preliminary informal evaluations, the latent space’s magnitude does not correlate with

Table 6.1: Pearson r and Spearman ρ correlation values for the corpus’s latent spaces and DFT phase distances.

Latent Space	DFT $\{0, 1\}^{12}$		DFT $\{0, 1\}^{128}$	
	Pearson (r)	Spearman (ρ)	Pearson (r)	Spearman (ρ)
Cosine	.65 $\pm$.04	.65 $\pm$.05	.64 $\pm$.04	.61 $\pm$.05
Euclidean	.61 $\pm$.06	.59 $\pm$.06	.65 $\pm$.05	.56 $\pm$.07

either DFT space, meaning that it is probably being used for non-tonal information. To inspect what information is captured, further research is required.

Secondly, the results reveal that the correlation between MusicVAE’s latent space is only residually higher with the DFT $\{0, 1\}^{12}$ space ($\rho \approx 0.65 \pm 0.05$) than the DFT $\{0, 1\}^{128}$ space ($\rho \approx 0.61 \pm 0.05$). According to Section 6.1, this might suggest that MusicVAE primarily identifies the latent ‘semantic’ meaning of harmony as the primary organizing principle, as harmonic function is primarily captured in DFT $\{0, 1\}^{12}$ space.

Even though MusicVAE is trained on a 128-bin piano roll representation, including octave and inversion information, the VAE’s bottleneck forces a prioritization of semantic invariants. In tonal music, the most powerful invariant is harmonic function. Thus, MusicVAE seems to learn underlying functional roles over specific register displacements, implying that it prioritizes abstract pitch-class relationships over literal pitch retention.

Nonetheless, these close correlation values still support interpreting chord distances as parsimonious voice leading. A hierarchical tension defines the space: harmonic function organizes it semantically, while the physical proximity of parsimonious voice leading influences it. In Bach’s style, these forces mutually reinforce each other, as the most efficient physical voice movements actively articulate the functional grammar of the chorales.

The nuanced distinction between these two spaces introduces the consideration of style, which is heavily influenced by the curated corpus used for training. Bach’s chorales are key examples of functional harmony, but they are also representative of parsimonious voice leading. In chorale music, the pitch range of the voices often mixes, which creates parsimonious horizontal voice-leading. Looking at a broader range of music (i.e., different instruments or a variety of composers) would help us better understand the differences between these two aspects. This comparison could shed light on how harmonic functions develop while also considering the unique challenges of leading voices, which vary depending on the musical style and time period.

6.3.1 Case Study: Chorale BWV 307

To further illustrate whether the latent space functions primarily as a common-tone functional space or a voice-leading space, we examine the opening phrase of J.S. Bach’s chorale *BWV 307* (B $\flat$ Major). This phrase provides an ideal case study because its functional harmony is typical of non-modulating passages, where the model exhibits its highest correlation values. Moreover, the chords in this passage are formed with concise voice leadings between individual voices.

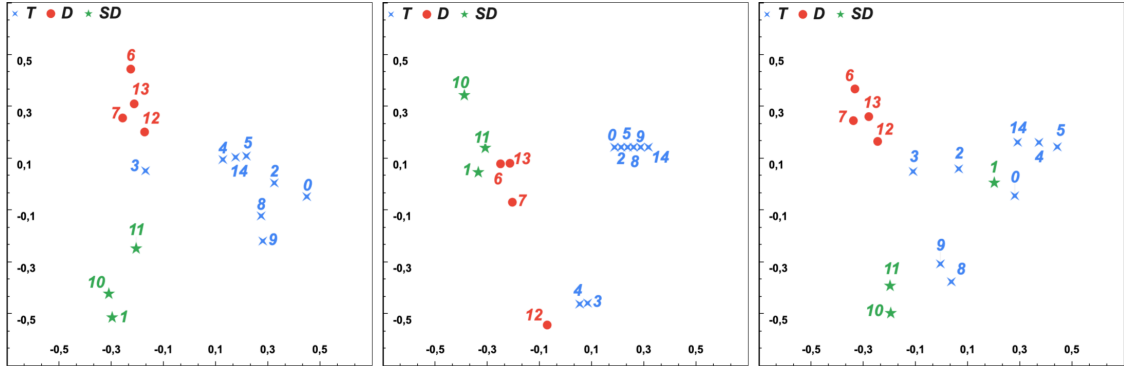


Figure 6.3: Comparative two-dimensional LLE visualizations of the latent space (left), the $\{0, 1\}^{12}$ pitch class DFT phase space (middle), and the $\{0, 1\}^{128}$ pitch distribution space (right) for the opening phrase of J.S. Bach’s chorale *BWV 307*. Tonal functional clusters are distinguished by color and marker type: Tonic (T) in $\times$, Dominant (D) in $\circ$, and Subdominant (SD) in $\star$. The plots illustrate how the latent space effectively preserves, and in certain instances clarifies, the functional organization found in theoretical common-tone phase spaces.

6.3.1.1 Geometric Visualization through LLE

Figure 6.3 showcases a two-dimensional visualization of the latent space alongside the DFT phase spaces for pitch classes and pitch distributions. We utilize Locally Linear Embedding (LLE) for dimensionality reduction. We chose this algorithm because it is known to preserve local structures and capture the intricate, nonlinear relationships inherent in tonal manifolds (Roweis and Saul, 2000).

In these visualizations, the tonic (T), dominant (D), and subdominant (SD) functions are distinguished by color and shape ($\times$, $\circ$, and $\star$, respectively). The latent space shows apparent differences in the harmonic function groupings. In fact, the way these functional groups are separated is even more straightforward than in the DFT $\{0, 1\}^{128}$ space. In the latter, for example, the initial subdominant chord blends into the tonic cluster. This indicates that the latent space might actually provide a more refined organization of functions compared to what we see in DFT theoretical spaces.

We believe that the way MusicVAE improves its understanding of music is because it works as a syntactic filter. While traditional DFT methods focus narrowly on the musical characteristics of individual chords, our model, based on a recurrent architecture, can learn broader patterns from the rich harmonic styles found in Bach’s music, like common progressions such as $I - ii - V - I$. As voice-leading choices can be more diverse per chorale than harmonic function, the structural learning process of MusicVAE seems to better generalize tonal functions. This implies that the model does not just replicate theoretical music concepts. Rather, it reshapes them into a clearer understanding of how tonal music works.

6.3.1.2 Sequential Proximity and Stylistic Influence

Figure 6.4 illustrates the chords of *BWV 307*, ordered by their proximity to the initial chord (chord 0). Even though the excerpt follows a parsimonious style of voice-leading, keeping one note stable from the previous chord while the others shift slightly, the bass note stylistically exhibits more significant jumps.



Figure 6.4: Analysis of the initial phrase of J.S. Bach’s chorale *BWV 307* (Bb Major). The musical sequences are organized as follows: (1) the original temporal sequence with functional harmonic analysis; and the chords ordered by proximity to the initial chord (chord 0) within (2) the VAE latent space, (3) the $\{0, 1\}^{12}$ pitch class DFT space, and (4) the $\{0, 1\}^{128}$ pitch distribution space. Red annotations indicate the number of common tones shared with chord 0, while boxed groups denote chords with equivalent distances to the reference.

In line with our corpus-wide findings, the sequential ordering indicates similar rankings throughout all spaces. This confirms that the distances in the latent space correspond with DFT hierarchical representations. Specifically, MusicVAE’s latent space aligns with a transpositional rotational space that preserves the circle of fifths and the relationships between relative major and minor tonal regional spaces.

On analysing the figure, we can perceive two intriguing examples where the latent space deviates from DFT theoretical expectation regarding voice-leading. The first example is the proximity of chords seven and eleven, likely due to the training data, where the transition from chord 0 to 7 occurs more frequently than from 0 to 11. The second example happens between chords one (*ii7*) and 13 (*V*). While voice-leading logic might suggest chord 1 is closer due to interval size, the latent space places chord 13 closer to the tonic. This aligns with the higher prevalence of *I – V* transitions in the Bach corpus compared to *I – ii7*.

6.3.1.3 Conclusion of the Case Study

The observations in this case study underscore a fundamental characteristic of MusicVAE: in the context of Bach chorales, the latent space tends to prioritize harmonic function over voice leading. Furthermore, because of the model's recurrent learning capability, it incorporates the statistical likelihood of harmonic progressions into its geometric distance. Consequently, the latent space represents a synthesis of theoretical ground truth and the specific compositional grammar of the training corpus.

6.4 Summary: Musicological and Creative Implications

The findings presented in this chapter demonstrate that the MusicVAE latent space transcends simple data compression, functioning instead as a geometrically meaningful representation of tonal logic (Carvalho and Bernardes, 2024b). The robust correlation coefficients establish a formal link between learned computational structures and established music theory.

The subtle discrepancy in correlation values, where the pitch class space ($\rho \approx .65$) outperforms the pitch distribution space ($\rho \approx .61$), reveals a crucial hierarchical trade-off in how generative models encode harmony. The VAE's optimization process prioritizes high-level tonal regions and functional clusters over granular voice-leading details. This suggests that the model is primarily semantically driven, favoring the organization of harmonic roles (Tonic, Dominant, Subdominant) as defined by pitch-class relationships.

These results provide a window into the stylistic characteristics embedded within the model. The deviations observed between the latent space rank and the theoretical DFT phase space are not necessarily computational errors. Rather, they highlight the specific harmonic and voice-leading conventions of the J.S. Bach chorale corpus. The latent space captures the transition frequencies and sequence of events unique to this style, prioritizing harmonic function in a manner that mirrors the compositional logic of the training data.

The ability of a VAE to internally derive structures that align with Fourier-based phase spaces opens new avenues for explainable AI in musicology. Future work involves exploring how these DFT phase spaces can be explicitly incorporated into the VAE's loss function as external regularization terms. This would transform the VAE from a purely data-driven model into a structurally hybrid system, guided by both statistical patterns and rigorous theoretical principles.

By exposing the rotational symmetries and topological continuities inherent in pitch-class distributions, we have established a framework where musical meaning is encoded as spatial distance. This provides the necessary foundation for the following chapter, which investigates how these high-dimensional maps, now proven to be in tune with theoretical phase spaces, can be navigated as interactive landscapes for both rigorous musicological analysis and real-time creative practice.

Chapter 7

Sampling Strategies for Musical Generation in Tonal Latent Spaces

The previous chapter presented empirical evidence of alignment between the MusicVAE’s latent space and DFT phase spaces. In this sense, comprehending that this latent space representation preserves the geometry of music theoretical hierarchies, our next question arises: “How can one navigate and sample these spaces to produce musically coherent and expressive results?”

This chapter examines different strategies of sampling music that help explore new ideas while still following traditional tonal musical rules. It focuses on finding a balance between creativity and structure, showing how specific methods can maintain predictable harmony (i.e., grammar, functional harmony, and parsimonious voice leading) while allowing for novel musical ideas. Traditionally, we can explore the smooth VAE latent space by using traditional linear interpolation. However, this approach has a drawback. While it is effective for connecting two musical segments (A and C), it can lead to paths that go through areas that produce musically incoherent or invalid results. To tackle this issue, we look at different sampling methods, including a discrete approach to linear interpolation, as well as discrete k -nearest neighbors and an angular-aligned version of k -nearest neighbors. Ultimately, the chapter aims to reveal how computers can balance creative diversity with musical coherence.

The chapter’s outline starts by defining navigation within latent spaces in Section 7.1. Then, Section 7.2 delineates the proposed sampling strategies that will be the basis for our experiment. Subsequently, in Section 7.3, we will describe the methodological approach for our experiment, which intends to evaluate the efficacy of these sampling strategies within the JSB corpus, adhering to stylistic and structural constraints of tonal harmony. Section 7.4 will present and examine the results of the experiment. At last, Section 7.5 concludes the chapter, offering a summary of our experiment and its principal implications to construct true interactive generative systems.

7.1 The Interpolation Problem in Latent Space

Computer music generation using deep learning models usually requires interpolating between two different existing musical ideas (Briot and Pachet, 2020). The process requires establishing a connection between their respective vectors (i.e., $\vec{A}$ and $\vec{C}$), which represent musical elements in latent space. Still, the learned manifold presents a challenge because it contains essential musical information that exists in a highly complex, non-linear framework. Within this learned space, the linear interpolation (LI) technique is the most usual approach, as it requires less computational power. However, its implementation usually results in artifacts because it needs to follow the musical rules that exist in the training material (Ji et al., 2023).

7.1.1 Sampling Strategies and Latent Space Topology

Outside of LI, the literature identifies several strategies for sampling VAE spaces to ensure generated materials remain musically coherent. For example, while LI works under the assumption of a straightforward geometric Euclidean space, in reality, the high-dimensional Gaussian distributions of VAE models often tend to gather their data points along the surface of a hypersphere (Jiménez and Landgrebe, 1998). Consequently, Spherical LI is frequently applied to navigate the shell of the distribution more effectively (White, 2016).

Furthermore, to fill up the space or generate variations, stochastic decoding (Yingzhen and Mandt, 2018) and Attribute Vector manipulation (Roberts et al., 2018) allow for targeted sampling, identifying specific directions in the underlying structure of music (e.g., rhythmic density or melodic contour). Using this information, we can intentionally choose points that meet specific musical or stylistic requirements, instead of just randomly moving from point $\vec{A}$ to point $\vec{C}$.

7.1.2 Continuous Space vs. Discrete Sampling

Departing from our experiments in Chapter 6, we employ a MusicVAE architecture, which utilizes a hierarchical recurrent decoder composed of a conductor RNN that breaks the latent vector into smaller, bar-level embeddings, which are then passed to a lower-level sub-decoder (Roberts et al., 2018). This structure enforces a global temporal coherence that a standard VAE lacks, ensuring that the interpolated middle chord (B') remains contextually grounded between A and C .

While the MusicVAE latent space is mathematically continuous, musical language (particularly the Bachian chorale) is fundamentally discrete. A raw coordinate sampled from the center of the manifold might decode into a hallucinated chord that lacks a clear functional identity or violates the strict voice-leading rules of the corpus. These regions are often referred to as latent dead zones where the model's generative confidence is low (Bryan-Kinns et al., 2021).

To address this, our experiment focuses on discrete sampling strategies. In this sense, instead of randomly picking a point along the linear path between two musical elements, we treat the latent space as a searchable dictionary of existing musical states. Employing this methodology ensures that the sampled points are connected to the real musical elements that the model learned

from, helping to balance the free-flowing nature of its creativity with the specific rules of music harmony.

7.2 Comparative Sampling Strategies

Taking into account the requirement for sampling discrete points within MusicVAE’s continuous manifold, we propose three primary strategies to navigate the latent space. Our main objective is to evaluate how different navigation geometries affect the stylistic fidelity of generated transitions, following the musical rules defined by Bach. We define this fidelity not as an abstract quality, but as the ability of the system to select a bridge chord (we will call it $\vec{B}'$) that keeps functional harmonic logic and parsimonious voice-leading consistent with the J.S. Bach corpus.

By testing these strategies on three-chord Bach sequences, we transition from theory to a practical challenge: figuring out which mathematical approach best navigates the musically incoherent zones within the latent space. We seek to understand how the geometry of movement (linear vs. directional) and the method of selection (distance-based vs. trajectory-based) affect the quality of the generated chord sequences. We have selected the following three strategies to represent a progression of solutions to the topological challenges of VAEs, which we will detail in the following sections:

1. **Linear Interpolation with discrete re-quantization (LI)**: This serves as our experimental baseline. It tests the naïve assumption that the latent space is Euclidean. By snapping a continuous linear midpoint to the nearest existing chord in the corpus, we can determine if simple geometric proximity is sufficient to preserve tonal function.
2. **k -Nearest Neighbors (k -NN)**: This strategy is motivated by the density-based nature of the learned manifold. Rather than following a geometric path, k -NN prioritizes the neighborhood of valid data points. This allows us to test if staying anchored to the model’s highest-confidence regions (the training data) outperforms the calculated paths of standard interpolation.
3. **k -Nearest Neighbors with Angular Displacement (k -NN+AD)**: This strategy is specifically designed to account for the hyperspherical topology of Gaussian latent spaces. Since high-dimensional data often clusters on a shell where direction is more meaningful than Euclidean distance, we use this method to see if preserving the “tonal momentum”¹ yields more stylistically natural substitutions. While Spherical LI is a standard navigation strategy within Gaussian manifolds, our focus on discrete re-quantization renders the specific curvature of the path secondary to the local density of the corpus. Therefore, we employ k -NN+AD as our representative strategy for capturing the manifold’s non-linear orientation.

¹Tonal momentum refers to the directional flow or angular trajectory between chords within a high-dimensional latent musical space. Rather than just measuring the raw spatial distance between two isolated musical elements, tonal momentum captures the concept that the geometric orientation of a vector in this space carries specific functional and musical meaning, such as a phrase’s forward movement toward a dominant chord or a cadential resolution.

7.2.1 Linear Interpolation (LI)

Conventional generative music methodologies using VAE latent spaces often entail sampling via linearly interpolated coordinates between two designated musical elements (i.e., chords, measures, or phrases), existing across different time scales (Pati et al., 2019; Prang and Esling, 2020; Ji et al., 2023). As our baseline control, linear interpolation (LI) tests the most direct path between two points in the n -dimensional space. For this method, we calculate a predetermined number of points at equal intervals between the starting vector $\vec{A}$ and the ending vector $\vec{C}$ using the equation:

$$\vec{B} = (1 - t) \cdot \vec{A} + t \cdot \vec{C} \quad , \quad (7.1)$$

where $t \in [0, 1]$ serves as the interpolation parameter. This parameter determines the weight of the destination point, aiming to facilitate smooth transitions based on the assumption of a continuous musical spectrum (Mi et al., 2021).

To bridge the gap between this continuous mathematical path and the discrete nature of Bachian harmony, we apply discrete re-quantization. For every calculated midpoint $\vec{B}$, we search the latent dictionary for the discrete chord point with the minimal Euclidean distance. This strategy is specifically chosen to quantify the straight-line failure hypothesized in Section 7.1.1. By forcing the interpolation to snap to real data points, we measure whether the shortest geometric path in latent space corresponds to a stylistically valid path in musical space, or if it instead drifts through functionally ambiguous latent dead zones (Bryan-Kinns et al., 2021).

Figure 7.1 illustrates this technique. Starting from points $\vec{A}$ and $\vec{C}$ within the latent space, a midpoint is generated by setting $t = 0.5$. The blue circles represent discrete chords projected into the latent space, while the dark-green hexagon represents the calculated linear midpoint. In this specific scenario, the nearest discrete point happens to be the original chord $\vec{B}$. Therefore, to explore novel variations, the sampled $\vec{B}'$ chord corresponds to the subsequent nearest neighbor in the discrete latent dictionary.

7.2.2 k -Nearest Neighbors (k -NN)

Moving beyond the baseline, the k -nearest neighbors (k -NN) algorithm addresses the structural reliability of the generation. Even though the MusicVAE latent space is mathematically continuous, music itself is made up of discrete elements. Instead of just taking the middle point between two elements as a potentially valid option, we see the latent space as a rich map where nearby elements show local musically-meaningful patterns that might be overlooked by global linear sampling strategies (Huron, 2001).

The k -NN strategy helps us find musical elements that are similar to each other based on these local patterns. The generated music maintains the existing styles of the dataset because the output only includes actual chords from the corpus, which creates authentic musical compositions that use real, functional, corpus-based chords. In the process, as shown in Figure 7.2, we calculate how similar a target music chord $\vec{B}$ is to all other chords in a multidimensional space. By finding the k closest chords, we can test if local data density is a better predictor of stylistic fidelity than raw

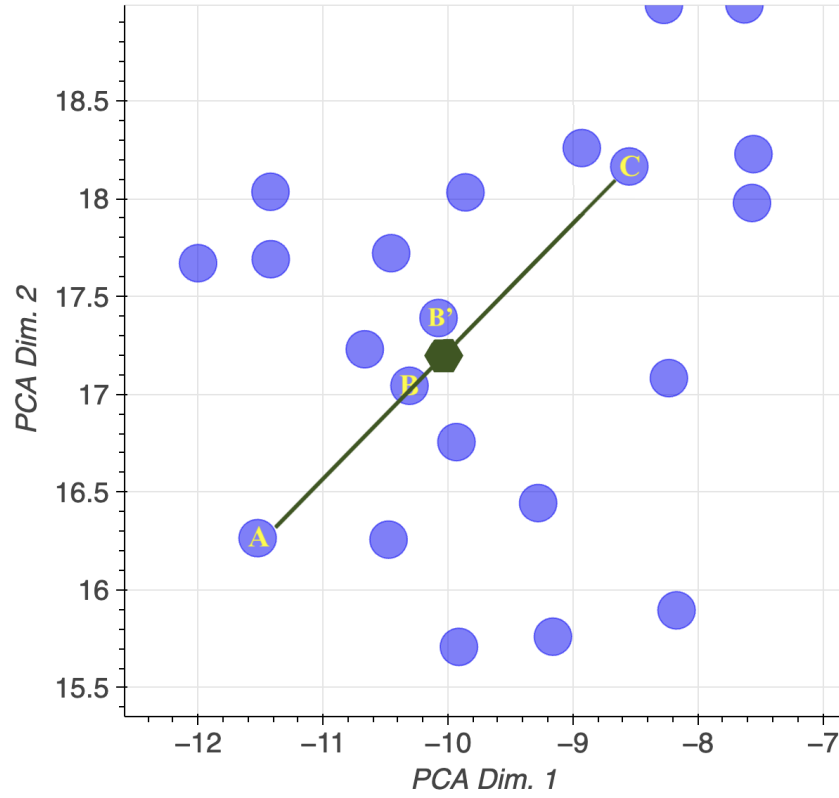


Figure 7.1: Sampling procedure for a synthetic chord $\vec{B}'$ using traditional linear interpolation between selected chords A and C. Discrete chords from the corpus are projected into the latent space (blue circles). The geometric midpoint (dark-green hexagon) is calculated as $\vec{z} = 0.5\vec{A} + 0.5\vec{C}$. Because this coordinate does not necessarily coincide with an existing musical element, the system samples the nearest available data point ($\vec{B}$) or, to increase variation, the second nearest neighbor ($\vec{B}'$).

geometric position. These neighbors serve as candidates for sampling, providing a diverse set of potential chords that are functionally and stylistically related to the target (Dasarathy, 1991).

As displayed in Figure 7.2, we use the vector $\vec{B}$ to establish a base point, which it will use to locate a replacement vector. The search process occurs within the area that extends from the center to the k nearest neighboring points, which the system identifies through its red dotted line (here it indicates a distance of three units, $k = 3$). The closest candidate according to the chosen distance metric is notated as $\vec{B}''$. With this strategy, we intend to preserve Bach chorale corpus styles by using latent space movements that connect adjacent points through their established harmonic and voice-leading boundaries.

7.2.3 k -NN with Angular Displacement (k -NN+AD)

The final strategy, k -NN+AD, is our most specialized approach, designed to preserve tonal momentum. This method is motivated by the topological nature of high-dimensional Gaussian manifolds. As recently demonstrated by Ascarate et al. (2025), the generative capacity of VAEs is

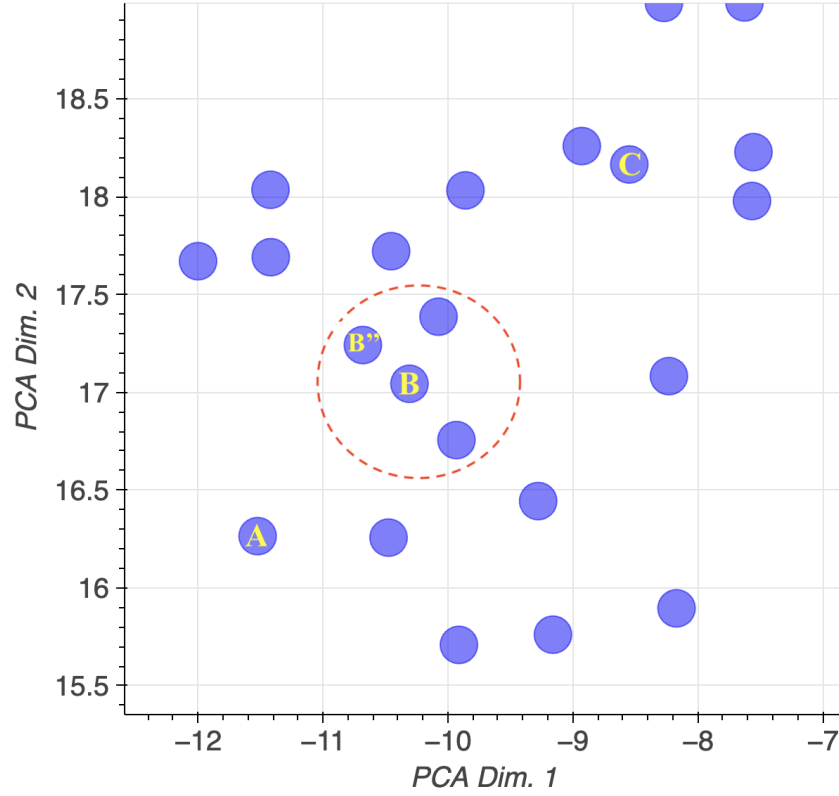


Figure 7.2: Sampling a substitute chord $\vec{B}''$ using k -nearest neighbors (k -NN). Discrete chords from the corpus are projected into the latent space (blue circles). For a given target vector $\vec{B}$, the k nearest neighbors (here, $k = 3$) are identified within the local neighborhood defined by the dotted red boundary. The sampled point $\vec{B}''$ is selected from this verified subset, ensuring the output remains anchored to valid data points while avoiding the low-density regions of the manifold.

significantly improved by utilizing hyperspherical coordinates, reinforcing the principle that in these spaces, the orientation of a vector is often more musically significant than its magnitude (Dong et al., 2019). By applying an angular filter to the sampling process, we test the hypothesis that maintaining the relative trajectory of a musical progression ensures more coherent functional resolutions.

To implement this strategy, we first select the k -nearest neighbors for our target vector $\vec{B}$ (usually, we select bigger k values than in the standard k -NN approach to broaden our selection pool). From this group, we prioritize the points that follow the same direction from the starting vector $\vec{A}$. Therefore, we create a reference vector, $\vec{AB}$, that represents the original musical transition and calculate the angular difference to the other candidate vectors that go from $\vec{A}$ to the potential neighbors. This similarity is measured using cosine similarity:

$$\text{similarity} = \cos(\theta) = \frac{\vec{AB} \cdot \vec{AB}_k}{\|\vec{AB}\| \|\vec{AB}_k\|} \quad (7.2)$$

As illustrated in Figure 7.3, k -NN+AD leverages the geometric flow of the latent path. A candidate point is selected as the optimal substitute as the optimal substitute ($\vec{B}'''$) if it exhibits the

minimal angular deviation. This happens even if it is not the absolute closest neighbor in terms of Euclidean distance. Thus, the mathematical direction of the latent path is translated into coherent musical transitioning, hopefully mirroring the common-tone and functional flow prevalent in the Bach chorale corpus.

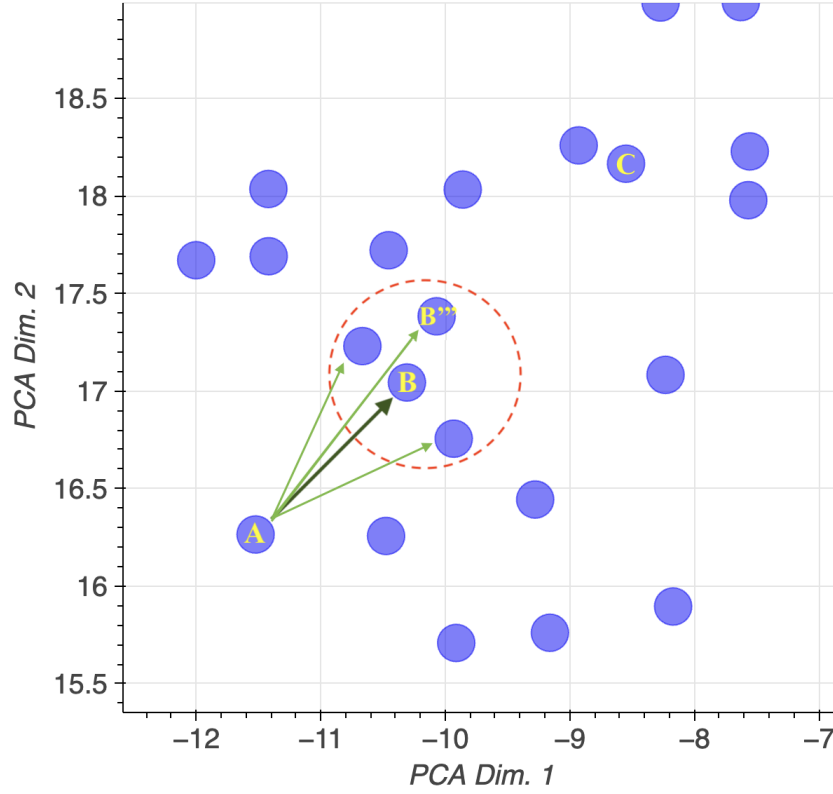


Figure 7.3: Sampling procedure for substitute $\vec{B}'''$ using k -NN+AD relative to the trajectory $[\vec{A}, \vec{B}]$. Discrete corpus chords are represented by blue circles, with the k nearest neighbors of $\vec{B}$ identified within the red dotted boundary. The dark-green arrow denotes the original latent vector $\vec{AB}$. The light-green arrows represent the vectors from $\vec{A}$ to each neighbor candidate. The selection of point $\vec{B}'''$ is attributed to the point whose vector exhibits the minimal angular deviation from the original path, prioritizing directional continuity over raw proximity.

This strategy builds on the phase-based organization discussed in Chapter 6. By focusing on the trajectory, we aim to keep the substitute chord in the same tonal direction as the original. This way, we respect the strict voice-leading and functional constraints of tonal harmony while also giving room for stylistic variation.

7.3 Evaluation Methodology

With the sampling strategies defined in Section 7.2, we are now able to detail the methodology of our experiment. The experiment seeks to determine the effectiveness of each sampling strategy to navigate the latent space, while respecting the stylistic and structural constraints of tonal harmony imposed by the corpus of Bach’s chorales.

The implementation code and supplementary data for the experiment have been made available online.² Furthermore, we provide an interactive platform³ that allows users to explore the high-dimensional space of the J.S. Bach chorale corpus in real-time. They can sample from the space, using chorale sequences as references, and see how the different strategies change the chord outcome. It connects complex math concepts with a more intuitive and hands-on experience.

The evaluation process is structured into two distinct phases. First, we use quantitative analysis to evaluate how accurately the generated trajectories match up with musical expectations. This means measuring how far the generated chords are from the original within a tonal phase space. With this evaluation, we want to make sure that the created chord sequences stay in musically coherent areas. The second phase of this evaluation consists of a qualitative musical analysis based on two case studies: 1) an authentic cadence in D Major (*BWV 184.5*) and 2) a half cadence in B Minor (*BWV 311*). The goal of this evaluation is to test, in critical musical contexts, whether these strategies capture key principles such as functional chord substitution and parsimonious voice leading. This dual approach allows us to verify if the latent space’s structural fluidity remains anchored in the underlying grammar of the Bachian style.

7.3.1 Materials: Model and Dataset

The foundational materials used for this experiment are the same ones introduced in the previous chapters: MusicVAE as the core VAE model and J.S. Bach chorales as the representative corpus of tonal functional harmony. Here, we summarize the specific parameters used for this evaluation.

To construct the latent space, we utilize MusicVAE. This architecture features a bidirectional LSTM encoder (two layers of 1024 units) and a hierarchical unidirectional LSTM decoder. A latent size of 256 is selected to balance dimensionality reduction with information density (Prang and Esling, 2020). The model is trained on binary 128-element salami slice piano roll encodings, capturing intricate pitch changes as they occur on the musical surface.

The stylistic consistency of the 371 J.S. Bach chorales (i.e., stable textures, limited pitch registers, and precise harmonic rhythms) makes them an ideal benchmark for studying latent harmonic organization. The corpus, obtained through the music21 library, features a nearly equal distribution of major (53%) and minor (47%) keys. For training, we utilize 60% of the chorales, augmented through twelve-tone transposition, while the remaining 40% are reserved for testing.

7.3.2 Experimental Hypothesis

The central research hypothesis of this study is that using density-based discrete methods to navigate the latent space will achieve better stylistic accuracy with Bach’s musical corpus than using continuous geometric methods. We established three specific testable predictions to define the sampling strategies’ fidelity:

²<https://github.com/NadiaCarvalho/SMC-BachLatentNavigation>, last accessed January 29, 2026.

³BachLatentNavigator is available at <https://nadiacarvalho.github.io/BachLatentNavigator/>, last accessed January 29, 2026.

- **H1: Mitigation of the Latent Dead Zone Effect:** We hypothesize that LI, even with re-quantization, will consistently produce the highest mean distance in the TIV-weighted pitch space (Eq. 7.4). This would confirm that a straight-line path bypasses functional clusters, resulting in chords that lack clear Bachian identity.
- **H2: Functional Substitution via Proximity:** We predict that k -NN will produce the lowest mean distance to the original corpus. This assumes that the latent space’s local topology accurately encodes harmonic functions, allowing for substitutions that maintain the same vertical sonority and structural role as the training data.
- **H3: Directional Preservation of Tonal Momentum:** We anticipate that k -NN+AD will be better at preserving the trajectory of a phrase. Specifically, while k -NN may optimize for raw chord similarity, k -NN+AD’s created transitions will better reflect directional voice-leading and functional resolutions (e.g., V-I cadential flows) of the original sequences.

7.3.3 Computational Method

To test these hypotheses, we conduct a large-scale comparative analysis across 28,600 test sequences (15,276 major and 13,324 minor). The experiment follows a three-step pipeline: 1) we extract original latent trajectories $[\vec{A}, \vec{B}, \vec{C}]$ from the test set; 2) for each sequence, we remove the middle chord $\vec{B}$ and generate a replacement $\vec{B}'$ using our three strategies (LI, k -NN, k -NN+AD); and 3) to quantify the musicality of the replacement, we project the resulting chords into a Tonal Interval Vector (TIV) space.

In the TIV framework, chords are represented as chroma vectors c_n and transformed into a weighted pitch-class DFT space using the equation:

$$T(q) = w_q \sum_{n=0}^{N-1} \tilde{c}_n e^{\frac{-j2\pi qn}{N}}, \quad 1 < q < 6 \quad (7.3)$$

where $N = 12$ and w_q are weights derived from empirical dissonance ratings (Almeida et al., 2022). This transformation is essential because it allows us to calculate the perceptual distance between the model’s output and Bach’s original intent. The stylistic fidelity is calculated as the Euclidean distance between the generated vector $\vec{T}(q)_{gen}$ and the original vector $\vec{T}(q)_{orig}$:

$$d\{T_1(q), T_2(q)\} = \sqrt{\sum_{q=1}^M |T_1(q) - T_2(q)|^2} \quad (7.4)$$

7.3.4 Result Interpretation

By analyzing the mean d values across the corpus, we provide a quantitative indicator of which sampling method most closely adheres to the underlying logic of J.S. Bach’s harmonic language. To provide context for the distances reported in Table 7.1, as the space is not normalized, we establish a set of reference distances within the TIV space for the key of C Major.

Within this framework, a distance of ≈ 4.5 represents a high-similarity common-tone substitution, such as the relationship between *C* major (*I*) and *A* minor (*vi*). A standard functional movement, such as the perfect fifth between *C* major (*I*) and *G* major (*V*), yields a distance of ≈ 9.0 , while a more dissonant diatonic relationship like *C* major (*I*) and *B* diminished (*viio*) results in a distance of ≈ 14.5 . Finally, a move to a tritone-distant key, such as *C* major (*I*) to *F#* major, produces a distance $\gtrapprox 22.0$.

These baseline values were calculated by projecting the chroma vectors of standard triads into the TIV space using the weighted DFT parameters w_q . Thus, the resulting Euclidean distances are a direct consequence of the common-tone overlap and the dissonance-based weights used to calibrate the manifold.

7.4 Results and Discussion

The comparative analysis of sampling strategies reveals significant differences in their ability to maintain the structural and stylistic integrity of the Bach chorale corpus. Table 7.1 summarizes the mean distances between the sampled candidates ($\vec{B}'$) and the original chords ($\vec{B}$) across major, minor, and combined datasets.

Table 7.1: Results of the comparative analysis regarding the perceptual relatedness of the chord sampling strategies across major, minor, and combined datasets.

Metric	Strategy	Major	Minor	All
Euclidean	Linear Interp.	19.6 ± 6.8	20.5 ± 6.8	20.1 ± 6.8
	<i>k</i> -NN	9.6 ± 4.3	10.2 ± 4.5	9.5 ± 4.3
	<i>k</i> -NN+AD	10.0 ± 4.6	10.7 ± 4.8	9.9 ± 4.5
Cosine	Linear Interp.	19.7 ± 7.2	20.6 ± 7.2	20.1 ± 7.2
	<i>k</i> -NN	9.5 ± 4.2	10.2 ± 4.4	9.4 ± 4.2
	<i>k</i> -NN+AD	10.0 ± 4.6	10.8 ± 4.8	9.8 ± 4.5

7.4.1 Quantitative Performance Analysis

The results presented in Table 7.1 provide strong empirical support for our primary hypotheses. Most notably, H1 is confirmed by the poor performance of LI. Across all datasets, LI yielded a mean Euclidean distance of ≈ 20.1 . When compared to our benchmarks in Section 7.3.4, this value approaches the distance of a tritone-distant chromatic chord ($d > 22.0$). This confirms that a straight-line path in the latent space often bypasses the functional clusters established in the training data, landing instead in musically neutral zones that lack specific functional identity within the Bachian style.

Similarly, the results strongly support our second hypothesis. The *k*-NN strategy consistently achieved the highest degree of perceptual similarity, with mean distances ($\rho \approx 9.5$) approximately

half those of the linear baseline. This value aligns almost perfectly with our benchmark for functional proximity ($I \rightarrow V$ at $d \approx 9.0$), indicating that k -NN successfully identifies candidates that reside within the same diatonic and functional neighborhood as the original intent.

The k -NN+AD strategy followed closely ($\rho \approx 9.9$). However, contrary to H3, the addition of angular constraints did not yield a statistically significant quantitative improvement in perceptual similarity over standard k -NN. While k -NN+AD respects the trajectory of the phrase, the raw proximity of k -NN appears to be a more reliable indicator of vertical harmonic validity.

7.4.2 Qualitative Musical Analysis: Case Studies

To evaluate whether these strategies capture key principles such as functional chord substitution and parsimonious voice leading, we examine two specific case studies.

7.4.2.1 Case Study 1: *BWV 184.5* (D Major Cadence)

In Figure 7.4, we examine an authentic cadence in D Major derived from J.S. Bach’s *BWV 184.5*. The figure displays the original sequence ($[\vec{A}, \vec{B}, \vec{C}]$) alongside the top five discrete chord candidates generated by each strategy to replace the middle dominant chord ($\vec{B}$). These candidates are ordered by their proximity to the target coordinate, providing a clear visualization of how each algorithm perceives the functional space.

Both k -NN and k -NN+AD strategies successfully identify candidates within the Dominant functional space, maintaining the original harmonic intent. While k -NN yielded a smoother bass line interpolation, k -NN+AD provided intriguing options for the upper voices, with the third and fourth candidates demonstrating resolution in opposing directions, a frequent attribute of Bach’s contrapuntal style.

In contrast, LI failed to maintain the cadential function. Because the original dominant chord is functionally distant from the surrounding tonic chords, the linear path pulled the sampled candidates back toward the tonic center. This resulted in transitions that lacked the necessary harmonic tension for a cadence, favoring common-tone retention at the expense of functional logic.

7.4.2.2 Case Study 2: *BWV 311* (B Minor Half Cadence)

A similar pattern emerged in the minor domain (Figure 7.5). For this half cadence ending on the dominant, LI again prioritized voice leading relative to the starting and ending points, thereby compromising the dominant function. The resulting chord was perceived as stylistically incoherent within the B-minor context, falling into the highly dissonant diatonic relationship range of our benchmarks.

While k -NN and k -NN+AD produced similar outcomes, k -NN was judged superior in this specific context. The k -NN+AD selection introduced a leap of a fourth in the upper voice, which, while directionally consistent in the latent space, proved less stylistically natural for a Bach chorale than the neighbor-based selection.

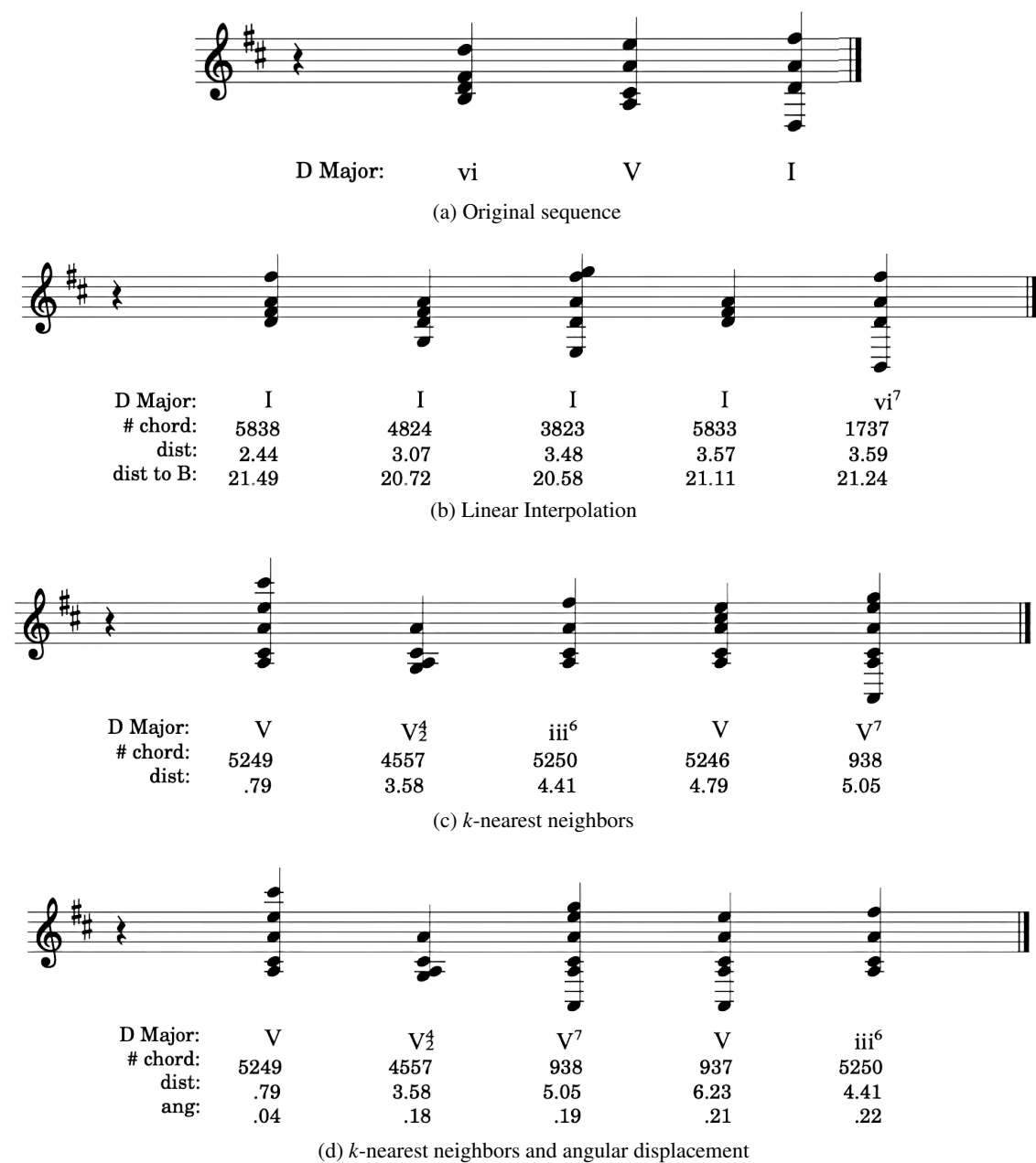


Figure 7.4: Original sequence, derived from the cadence of the first phrase of J.S. Bach’s Chorale *BWV 184.5* in D Major, along with five optimal sampling candidates using the three different sampling strategies, ordered by closeness to the original. For each strategy, we present the tonal functions within the phrase, the number of the chord in the dataset of all chords, and the distance metric calculated, i.e., in k -NN and k -NN+AD, we present the Euclidean distance (dist) from the candidate point to the original, plus angular distance for k -NN+AD. For LI, we present the Euclidean distance from the candidate point to the interpolated continuous point in the latent space (dist) and the distance to the original chord (dist to $\vec{B}$).

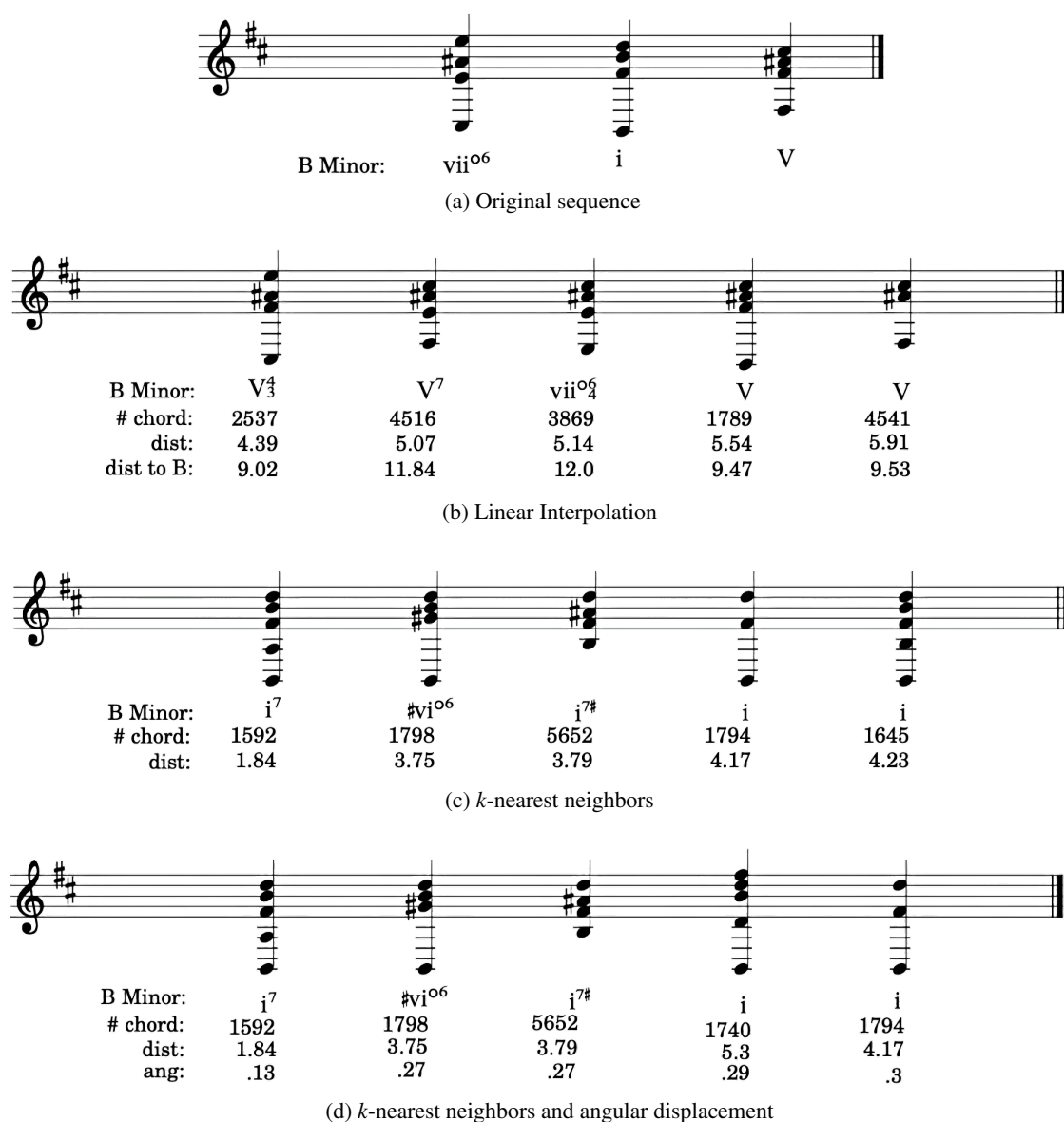


Figure 7.5: Original sequence, derived from the cadence of the first phrase of J.S. Bach’s Chorale *BWV 311* in B Minor, along with five optimal sampling candidates using the three different sampling strategies, ordered by closeness to the original. For each strategy, we present the tonal functions within the phrase, the number of the chord in the dataset of all chords, and the distance metric calculated, i.e., in k -NN and k -NN+AD, we present the Euclidean distance (dist) from the candidate point to the original, plus angular distance for k -NN with angular displacement (ang). For linear interpolation, we present both the Euclidean distance from the candidate point to the interpolated continuous point in the latent space (dist) and the distance to the original chord (dist to $\vec{B}$).

7.4.3 Synthesis: Functional Priority vs. Geometric Pathing

The failure of linear interpolation confirms that the MusicVAE latent space is non-linear with respect to functional harmony. A straight line between two tonics does not necessarily pass through

the dominant, even if a dominant chord is the musical midpoint of the sequence.

Furthermore, while angular displacement was intended to enhance voice-leading continuity, the results suggest that the latent space’s existing clustering (as captured by standard k -NN) already encodes much of the necessary stylistic information. The latent space thus functions as a semantic map where proximity inherently accounts for both functional role and, to a significant extent, the parsimonious voice-leading patterns prevalent in the training corpus. k -NN proves to be the most robust strategy for maintaining the grammatical integrity of the music. At the same time, k -NN+AD offers a specialized alternative for cases where the directional momentum of the phrase is the primary compositional concern.

7.5 Summary: Implications for Interactive Generative Systems

The experimental results presented in this chapter confirm that k -NN and k -NN+AD are better strategies for transitioning from abstract latent space structures to musically meaningful output. Our analysis demonstrates that LI is inadequate for navigating complex tonal manifolds. By failing to respect the non-linear functional clusters of the latent space, LI frequently produces transitions that are harmonically stagnant or stylistically inconsistent (Carvalho and Bernardes, 2024a).

The distinct performance characteristics of k -NN versus k -NN+AD provide a foundational design principle for future co-creative tools. While k -NN proved superior in maintaining the local continuity of the bass line and parsimonious voice-leading, k -NN+AD demonstrated a unique capacity for preserving the directional tonal momentum of a phrase. Accordingly, both strategies ensure that the overarching harmonic function of the chords remains stable even when introducing variation.

For the developer of interactive generative systems, these findings suggest the implementation of a hybrid control scheme. Rather than relying on a single navigation algorithm, a composer-centric interface could offer a specialized blending mechanism. In this framework, k -NN serves as the primary mode for voice-leading focus, proving ideal for tasks requiring smooth, step-wise transitions and textural consistency. Alternatively, k -NN+AD provides a robust harmonic-function focus, supporting generative tasks that demand strong functional logic and precise directional movement through the circle of fifths.

By providing such fine-grained, musically intuitive control, we move away from treating the VAE as an opaque black box and instead transform the latent space into a flexible, interactive canvas for guided musical creation. Consequently, having established the theoretical alignment of the latent space (Chapters 5 and 6) and the optimal strategies for navigating its terrain (Chapter 7), the research now shifts from analysis to implementation. The following chapter details the development of a topological instrument designed to put these findings into practice, exploring how these hybrid navigation strategies are integrated into a real-time interface that allows composers to explore the Bach-style manifold through gestural and semantic control interactively.

Chapter 8

BroadcastJSB: Navigating Disentangled Harmonic Spaces

This chapter details the design and implementation of BroadcastJSB,¹ a web-based interactive application that translates the theoretical findings of the preceding chapters into a creative tool.

The motivation for developing BroadcastJSB stems from a fundamental tension in contemporary AI-based generative music. As established in the Chapter 3, while deep learning models offer immense generative power, their multidimensional neural connections often introduce profound opacity into the creative process. This creates a black box effect that distances the composer from the underlying structural logic, resulting in a “control gap” where interaction relies on abstract statistical matching rather than intentional musical foresight. To overcome this, generative models must be transformed from unpredictable black box generators into transparent, navigable interfaces.

This application builds upon foundational efforts to create smart tools for harmonic content transformation, such as the HarT system (Cocharro and Bernardes, 2017), which demonstrated the value of allowing users to intuitively recreate the harmonic surface of a composition while strictly retaining its underlying rhythmic and tonal function structure. By merging knowledge-based rules with learning-from-example approaches, tools like HarT address the lack of long-term structure in generative music. BroadcastJSB extends this philosophy into the realm of deep learning, utilizing the topological geometry of the latent space itself as the site of interaction.

Through Chapters 5 and 6, we demonstrated that variational autoencoders (VAEs), particularly when utilizing spectral representations like the Discrete Fourier Transform of pitch (DFT-P), naturally organize their latent spaces to reflect cognitive tonal pitch spaces and theoretical Fourier phase spaces. However, as Chapter 7 revealed, navigating this continuous mathematical space poses a distinct performative challenge: simple linear interpolations frequently become trapped in “latent dead zones,” generating functionally ambiguous or musically incoherent artifacts. BroadcastJSB was developed to explicitly solve this interpolation problem by operationalizing intent-aware sampling strategies, specifically k -nearest neighbors (k -NN) and k -NN with

¹<https://nadiacarvalho.github.io/BroadcastJSB/>, last accessed January 29, 2026.

angular displacement (k -NN+AD), allowing for the retrieval of discrete, stylistically valid chords that preserve the formal voice-leading and “tonal momentum” of the original corpus.

Ultimately, the creative purpose of BroadcastJSB is to realize the musical work as a mutable interface. By leveraging the metaphor of an analog latent space radio receiver, the system implements Explainable AI (XAI) at the interface level, bridging the gap between abstract mathematical manifolds and the tactile experience of musical discovery. It deliberately defamiliarizes the J.S. Bach canon, shifting the creative experience from passively “listening to a score” to actively “exploring a style.” By adjusting a tuning dial to introduce calculated “drift,”² the user is empowered to navigate validated functional clusters, discover geometric transitions, or creatively extrapolate the harmonic trajectory of the chorales into highly chromatic, high-entropy territories. In doing so, BroadcastJSB transforms the J.S. Bach corpus into a fluid, navigable territory, where the machine acts as a legible, symbiotic partner in the real-time exploration of common-practice tonality.

The remainder of the chapter is structured as follows. Section 8.1 introduces the core concept of BroadcastJSB as a latent space radio receiver, where users manipulate a tuning dial to introduce drift and dynamically re-harmonize the Bach canon. Section 8.2 details the system’s bifurcated architecture, bridging offline manifold encoding with a highly responsive real-time web pipeline. Next, Section 8.3 outlines the three decoder strategies that operationalize user agency: the NEAR (k -NN) mode for localized variations, the MIDPOINT (linear interpolation, LI) mode for geometric transitions, and the VECTOR (k -NN+AD) mode to extrapolate tonal momentum. Finally, Sections 8.4 and 8.5 evaluate the system, demonstrating how analog-style visual and auditory interference act as a transparency layer to communicate algorithmic uncertainty, successfully transforming a black box generative model into an explainable, co-creative musical environment.

8.1 BroadcastJSB: The latent space Radio Receiver

The primary objective of BroadcastJSB is to provide a topological instrument that respects the cognitive and theoretical distances of Western common-practice harmony learned from a representative corpus of J.S. Bach chorales. While previous chapters established the latent space as a static map of harmonic relationships, BroadcastJSB transforms this mathematical geometry into a dynamic, exploratory environment. Ultimately, this transitions the system beyond the black box nature of generative AI, offering a mutable interface where the user can actively navigate and reharmonize the Bach canon.

8.1.1 The Radio Metaphor as a Cognitive Bridge

BroadcastJSB, whose web interface is depicted in Figure 8.1, is conceptualized as a latent space radio receiver that treats the high-dimensional vector space derived from the MusicVAE’s chord dictionary as a radio landscape. The metaphor of a radio receiver, complete with tuning dials,

²We define drift within this context as the controlled deviation from the composition’s original structural coordinates, which translates musically into a dynamic process of harmonic variation and re-harmonization of the Bach canon.

signal interference, and radio “stations,” is used to bridge the gap between abstract mathematical manifolds and the experience of musical discovery.



Figure 8.1: BroadcastJSB’s user interface. The design utilizes a mid-century radio metaphor to facilitate the exploration of the Bach chorale latent space. Key interactive elements include: the frequency window (top), which visualizes the drift from the original chords; the station window (center), showcasing the current chorale being played; the tuning dial (bottom center); and the navigation strategy switch (bottom right), which allows users to toggle between NEAR (k -NN), MIDPOINT (LI), and VECTOR (k -NN+AD) sampling strategies.

In this framework, the J.S. Bach corpus represents the broadcast signal, a state of maximum structural coherence. As the user moves away from the original coordinates, they are essentially “detuning” the receiver. In practical, musical terms, this detuning acts as a dynamic process of chord substitution. Because the latent space organizes chords by their theoretical and cognitive similarity, as established in Chapters 5 and 6, navigating away from the original chord introduces variations that structurally maintain their underlying harmonic function. This translates the ab-

tract black box of the generative model into an explainable, intelligible space: just as a radio dial navigates a linear frequency spectrum, the BroadcastJSB interface navigates a tonal spectrum where spatial distance directly corresponds to the degree of harmonic substitution. To illustrate the practical results of this explainable substitution, Section 8.4.1 presents comparative musical scores demonstrating the original Bach inputs alongside the resulting deviations at varying levels of interference.

8.1.2 Dimensionality and Agency

The transition from the 256-dimensional (256-D) latent space (discussed in Chapter 5) to a simple user interface (UI) requires a reduction of complexity that preserves the semantic integrity of the music. BroadcastJSB employs the 371 Bach chorales as its primary corpus, where every unique vertical aggregate (chord) is a discrete point in this space.

By turning a physical or virtual tuning dial, the user introduces drift into the harmonic progression. This mechanism parallels recent interactive systems by Banar et al. (2023), which demonstrate the value of using regularized latent spaces to generate controllable and explainable variations of musical themes. This drift is not a random perturbation but a calculated movement along the geodesics of the manifold that acts as a real-time process of functional chord substitution. Drawing from the sampling strategies analyzed in Chapter 7, this dial mediates the user's agency over the musical surface by altering the underlying voice-leading and harmonic syntax:

- **Ground Truth (0% Tuning):** The system plays the pure, intended harmony of Bach. The receiver is locked onto the station, reproducing the exact diatonic framework, parsimonious voice-leading, and functional harmonic logic of the original score.
- **Latent Variation (10-90% Tuning):** While the original chorale structure acts as a rhythmic scaffold, the system introduces a dynamic re-harmonization driven by common-tone proximity and tonal momentum. Depending on the active decoder strategy, the original chords are replaced by their closest functional neighbors (e.g., substituting a tonic *I* with a submediant *vi* to maintain high common-tone overlap), geometric midpoints, or directionally aligned vectors that preserve the phrase's cadential trajectory. Consequently, the generated music retains the grammatical essence of common-practice tonality while actively exploring novel diatonic and chromatic substitutions.
- **Entropy and Abstraction (100% Tuning):** The structural connection to the original score is eliminated. The harmonic progression transitions into high-entropy, highly chromatic regions of the latent space. In these areas, the decoder generates sonorities (such as tritone-distant or heavily altered chords) that are statistically plausible within the VAE's abstract geometry but functionally ambiguous, effectively bypassing traditional voice-leading resolutions and the original tonal center.

This approach fulfills the thesis's overarching established theme: the musical work as a mutable interface. By representing the Bach chorales not as fixed sequences of notes but as paths

through a latent space, BroadcastJSB allows for a non-linear engagement with the Western common-practice representative corpus of Bach chorales. Shifting from a passive listener to an active explorer, one does not just listen to Bach. One navigates the possibility space of Bach, using the mathematical alignment of the VAE (Chapter 6) to ensure that every station they land on maintains a degree of cognitive and musicological validity.

8.2 System Architecture

The architecture of BroadcastJSB is designed to handle the high computational load of real-time latent space navigation within a standard web browser. The system follows a modular pipeline, depicted in Figure 8.2, where high-level user gestures are translated into latent coordinates, queried against a multidimensional dictionary, and finally decoded into an audible musical surface.

8.2.1 Modular Pipeline: From Gesture to Sound

The architecture of BroadcastJSB is organized into a bifurcated pipeline. It distinguishes between an offline computational stage (responsible for manifold generation) and a real-time reactive stage (responsible for manifold navigation). This separation ensures that high-dimensional complexity does not compromise the synchronous requirements of live audio synthesis. The integrated pipeline, depicted in Figure 8.2, consists of four discrete stages:

1. **Manifold Encoding (Offline Pre-processing):** Using a music21-based Python pipeline, the source corpus (e.g., J.S. Bach’s 371 chorales) is vertically segmented into rhythmic slices. Each unique chord is encoded and mapped to a 256-D piano roll vector using the MusicVAE encoder. The result is a static Latent Dictionary and a set of Bach Chorale encoded sequences, providing the spatial and temporal maps for the receiver.
2. **Gestural Mapping (Interaction Capture):** At runtime, the user manipulates the tuning dial, which is bound to a reactive `tunerValue` in the Vue 3 frontend. This value is used to calculate the current radius of drift, ensuring that the high-dimensional displacement in the 256-D MusicVAE space feels perceptually intuitive to the user, providing fine-tuning control near the original Bach coordinates.
3. **Topological Query (Spatial Search):** The calculated drift parameters are dispatched to the broadcaster logic module. Here, the system performs an $O(\log N)$ query against the k -d tree dictionary. Depending on the active Decoder Strategy, be it the stochastic proximity of NEAR, the geometric averaging of MIDPOINT, or the trajectory-aware momentum of VECTOR, the search engine identifies the target pitch-class set that best satisfies the current topological constraints.
4. **Symbolic-to-Signal Resynthesis (Tone.js):** The identified chord ID is mapped back to its corresponding MIDI pitch-class values and dispatched to a polyphonic additive synthesizer,

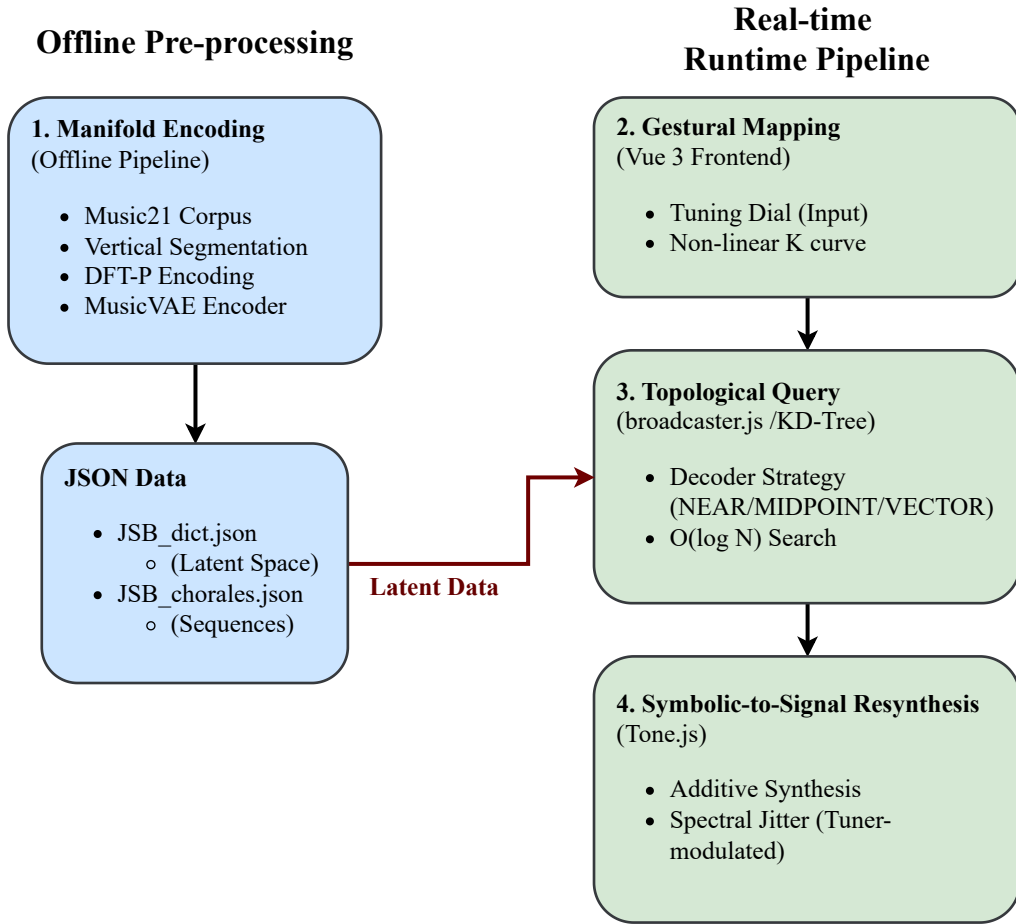


Figure 8.2: The bifurcated pipeline architecture of BroadcastJSB. The system decouples the Offline Pre-processing stage (left) from the Real-time Runtime Pipeline (right). During the former stage, the musical corpus is vertically segmented, encoded, and mapped to the MusicVAE latent space. In the runtime environment, user gestural inputs are mapped through a non-linear k -curve to drive an $O(\log N)$ spatial search via a k -d tree. The identified coordinates are then resynthesized in the signal domain using a tuner-modulated additive synthesis engine. This modularity ensures that the interface remains mutable, allowing for the swap of underlying musical corpora without altering the fundamental interaction logic or audio engine.

effectively bridging the symbolic and signal domains. While the latent query dictates the pitch content, the spectral cleanliness (e.g., noise floor, oscillator jitter, and filter cutoff) is modulated by the same `tunerValue`. This process generates a direct auditory representation of the mathematical displacement from the Bachian ground truth.

By treating these computational elements as discrete modules, BroadcastJSB establishes a mutable relationship with its underlying data. The runtime environment is designed to remain agnostic to the specific harmonic language encoded within the latent dictionary. Consequently, the foundational J.S. Bach corpus can be seamlessly swapped for alternative musical datasets, such as jazz standards or folk melodies, simply by replacing the JSON data structures generated during the Manifold Encoding stage. Because the broadcaster logic, the k -d tree spatial search, and the

additive resynthesis engine operate strictly on generalized vector distances rather than hard-coded musicological rules, the application operates as a generalized framework for interactive latent exploration. Ultimately, this modular architecture fulfills the thesis’s overarching goal: moving beyond opaque black box generative models to provide a universal topological instrument capable of navigating any encoded musical style.

The implementation of the BroadcastJSB system, including the pre-processing pipeline and the real-time latent navigation logic, is documented and available in a public repository.³

8.2.2 Frontend and Interaction (Vue 3 & High-Speed Reactivity)

To establish a highly responsive, single-page interface that functions as a radio dashboard, the frontend architecture of BroadcastJSB utilizes Vue 3⁴. The application implements a modular logic pattern (see Figure 8.3), which explicitly decouples the reactive UI and static offline datasets from the core computational engines. Ultimately, this strict separation of concerns guarantees that the high-dimensional complexity of the latent space does not compromise the synchronous demands of real-time audio resynthesis.

8.2.2.1 Computational Logic Engines

The architecture of the computational engines is partitioned into four distinct logic modules, ensuring that a robust computational backbone supports the radio receiver metaphor:

- **The Broadcaster (`broadcaster.js`):** This module centralizes the application state and manages the temporal synchronization between the original Bach chorale sequences and the generative drift. It functions as the master controller, orchestrating the flow between the user’s tuning gesture and the resulting musical output.
- **The Signal Chain (`audioPlayback.js`):** Utilizing Tone.js, this module manages the additive synthesis and the specific signal-to-noise ratio. Rather than simply triggering static MIDI events, it dynamically reconstructs chordal sonorities using multiple phase-aligned oscillators. It translates the symbolic pitch-class sets into audio, introducing the spectral artifacts that simulate analog interference.
- **The Navigator (`latentStrategies.js`):** This module serves as the implementation site for the k -d tree search algorithm. It abstracts the complexity of high-dimensional searching, providing a streamlined interface for the rest of the application to request neighbors or midpoints from the latent dictionary.
- **The Metric Engine (`distance.js`):** This module implements the dual-metric approach validated in Chapter 7. It calculates Euclidean distance for k -NN proximity and cosine similarity (k -NN+AD) to determine tonal momentum. These metrics allow the system to

³<https://github.com/NadiaCarvalho/BroadcastJSB/>, last accessed January 29, 2026.

⁴<https://vuejs.org>, last accessed January 29, 2026.

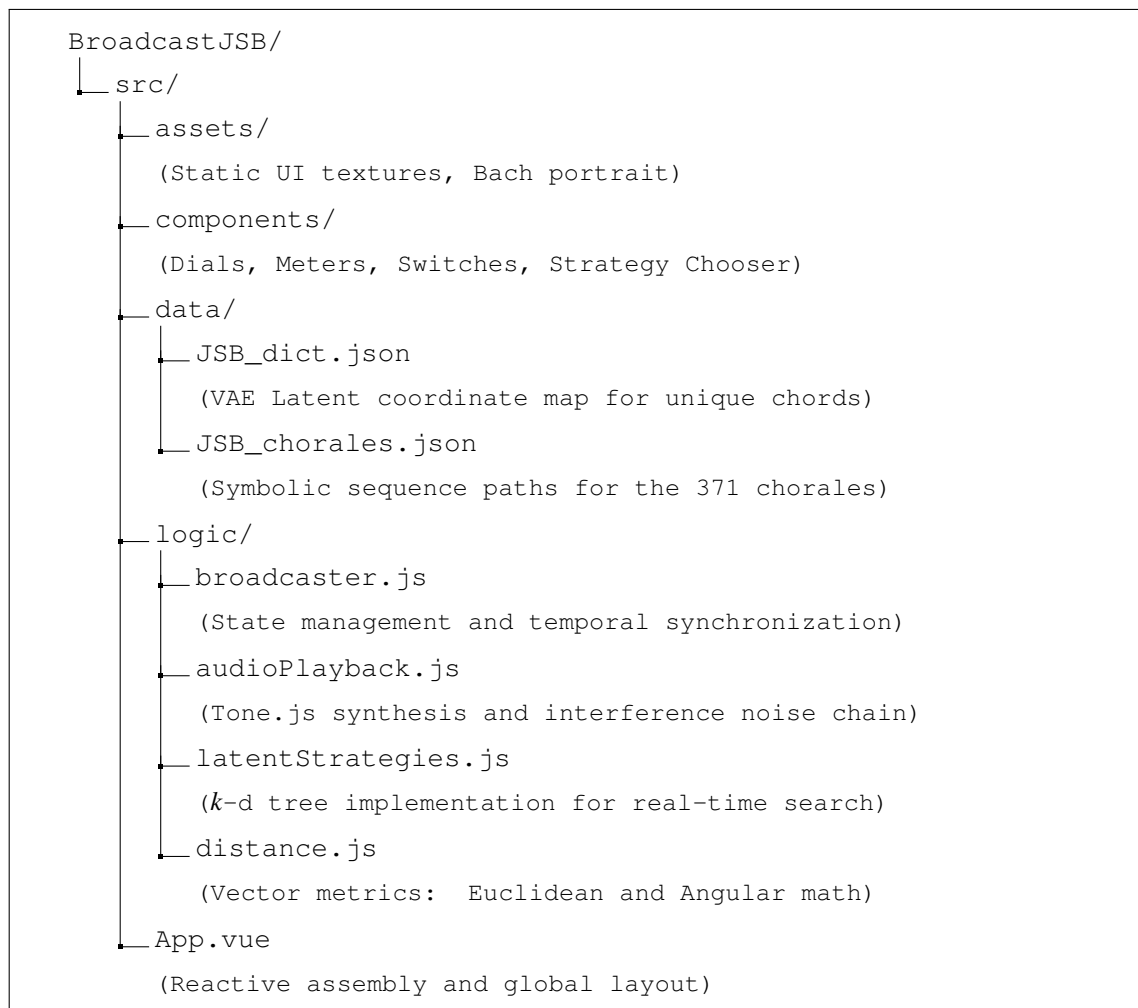


Figure 8.3: Frontend project structure of the BroadcastJSB application. The frontend architecture of BroadcastJSB follows a modular logic pattern that explicitly decouples the reactive Vue 3 user interface (`App.vue` and `components/`) and static offline datasets (`data/`) from the underlying computational engines (`logic/`). This separation ensures that complex state management, spatial searching via k -d trees, and vector mathematics do not compromise the synchronous requirements of the real-time audio synthesis.

differentiate between a neighbor that is near the target chord and a neighbor that is moving along the same direction within the 256-D manifold, ensuring that algorithmic substitutions strictly respect the directional voice-leading and functional logic of the Bachian style.

8.2.2.2 Spatial Search: Solving the Latent Proximity Problem (k -d tree)

A primary technical challenge identified when running the experiments in Chapter 7 is how to find the k most relevant chords among thousands of vectors in a 256-D space without taking too much time, which introduces audio latency. Relying on a standard brute-force search algorithm would require $O(N)$ computational operations for every single rhythmic salami slice of a chorale, a process that is far too slow for real-time generative audio.

To resolve this problem, BroadcastJSB employs a k -d tree (k -dimensional tree) data structure to manage the spatial search, using `kd-tree-javascript`.⁵ Instead of calculating the distance to every single point in the database, the k -d tree algorithm partitions the high-dimensional latent space into a hierarchical series of hyper-rectangles, effectively reducing the search complexity from $O(N)$ to $O(\log N)$.

In practice, this structural optimization allows the system’s navigator module to query the more than 2500 unique chord vectors within the Bach latent dictionary in under 2 milliseconds. Achieving this ultra-low latency is critical for the application’s performative viability. It guarantees that the computationally intensive spatial sorting does not bottleneck the additive resynthesis engine, thereby facilitating seamless, glitch-free musical transitions even as the user rapidly manipulates the tuning dial to drift between harmonic stations.

8.2.2.3 Audio Engine: Resynthesizing the Manifold (Tone.js & Web Audio)

Rather than relying on pre-recorded samples, which would impose a discrete limitation on the fluidity of the detuning effect, BroadcastJSB utilizes the `Tone.js`⁶ framework to implement a real-time additive synthesis engine. This choice allows the system to operate at the intersection of symbolic music representation and raw signal processing.

Upon identifying a target coordinate within the latent space (whether the ground-truth Bach chord or a chord substitute), the broadcaster module ensures the audio engine avoids merely triggering a static MIDI event. Instead, the system dynamically reconstructs the chordal sonority by layering multiple phase-aligned oscillators. Consequently, this architectural decision allows the mathematical distance from the original signal to be articulated directly through spectral transformations.

These spectral transformations manifest across three distinct operational states. For high-fidelity harmonic reconstruction, when the system is perfectly tuned (0% drift), it strictly prioritizes harmonic clarity. The oscillators are configured to produce clean, phase-aligned sine and triangle waves, replicating the purity and structural coherence of the Bachian broadcast signal. As drift is introduced via the tuning dial, the system generates latent-modulated spectral noise by calculating the Euclidean distance between the current selection and the ground truth. This distance is utilized as a control voltage for a noise-injection circuit, prompting the audio engine to inject filtered white noise and introduce microtonal jitter (frequency detuning) between the four voices. Ultimately, this creates an auditory representation of topology. The signal interference is not purely aesthetic. Rather, it serves as a tactile, auditory feedback mechanism where the detuning effect provides a psychoacoustic representation of the latent space’s density. As the system moves into noisier, less-populated regions of the manifold, the audio signal physically degrades, simulating the loss of a radio lock in the analog domain.

⁵<https://github.com/DefinitelyTyped/DefinitelyTyped/tree/master/types/kd-tree-javascript>, last accessed January 29, 2026.

⁶<https://tonejs.github.io>, last accessed January 29, 2026.

8.2.2.4 UI Design: The Analog-Industrial Interface

The visual layer employs textures and static assets (such as the Bach portrait and radio-dial textures) to anchor the advanced computational backend within a familiar Braun-inspired aesthetic. The design operates through a purely state-driven architecture, where the analog needle and flickering lights function as direct graphical representations of the raw z -coordinate offset and the distance metrics calculated in `distance.js`.

The interface is assembled in `App.vue` using specialized UI units located within the `components/` directory. These elements (i.e., dials, meters, and switches) are strictly designed to be dumb components. They receive their data via props from the `broadcaster.js` state and emit events when manipulated. Components such as the meters and the tuning dial leverage Vue’s reactivity to provide instantaneous visual feedback. For example, the meters are driven by the amplitude data from `audioPlayback.js`, establishing a direct sensory link between the latent coordinate and the visual signal strength. This streamlined, modular architecture allows BroadcastJSB to meet the real-time constraints required for a performative musical instrument while remaining true to the music-theoretical findings established in previous chapters.



(a) Latent proximity indicator in peak activation state.



(b) Ground-truth indicator with flicker artifacts (on/off).

Figure 8.4: Visual feedback mechanisms for latent signal tracking. (a) Latent proximity indicator registering a peak activation state to confirm coordinate alignment with the ground-truth chord (*B*). (b) Ground-truth indicator exhibiting visual fragmentation artifacts (flicker) during state transitions. Together, these elements translate the calculation of new latent trajectories into legible interface states.

The design of the interface utilizes reactive visual feedback to communicate coordinate alignment. A central feature is the latent proximity indicator (see Figure 8.4a), a graphical meter that visualizes the distance to the ground truth. When the system’s sampling logic successfully aligns with the original corpus coordinate, even amidst a high- k neighborhood, the meter registers a peak activation state (the meter’s needle glows in green). This visual confirmation ensures the user remains cognitively anchored to the source corpus.

Similarly, a ground-truth indicator, illustrated by a seal with Bach’s portrait (see Figure 8.4b), is tied to the state transitions between sampled units. The indicator exhibits visual fragmentation artifacts (“flickers”) during chorale transitions, representing the computational process of calculating a new latent trajectory. This indicator intends to mimic a radio’s physical act of “finding a frequency.”

8.2.3 Data Pipeline: From Music21 to JSON Dictionary

The underlying data that fuels the receiver is prepared through an offline Python pipeline⁷ designed to decouple harmonic identity from temporal structure. Using the music21⁸ library, the 371 Bach chorales were salami-sliced to extract every unique vertical chord aggregate (see Section 6.2.2).

Following this extraction, the aggregates are encoded via the piano roll format and processed through the MusicVAE encoder to generate their corresponding 256-D latent representations, as detailed in Chapter 6. Ultimately, to optimize real-time performance, the resulting dataset is partitioned into two distinct JSON structures:

- **The Latent Dictionary (`JSB_dict.json`):** Functioning as an atemporal map of the Bachian harmonic universe, this structure stores unique pitch-class sets indexed by their high-dimensional coordinates. Ultimately, this map acts as the search space for the LI, k -NN, and k -NN+AD calculations.
- **The Phrase Sequences (`JSB_chorales.json`):** Operating as a temporal database, this file organizes each of the 371 chorales into an ordered sequence of pointers. Each entry within the sequence contains an ID referencing a specific dictionary coordinate alongside a defined `duration` (measured in quarter-lengths).

Fundamentally, this bifurcated architecture empowers BroadcastJSB to maintain a robust rhythmic scaffold. By loading the temporal sequence data directly into the browser’s memory, the system perfectly preserves the characteristic metric pulse and voice-leading timing of the original Bach score. At the same time, the central broadcaster module retains complete flexibility to dynamically substitute the underlying harmonic content at every discrete chord slice. Consequently, this structural separation ensures that any applied drift is perceptually experienced as a complex re-harmonization of a familiar rhythmic gesture, rather than a complete dissolution of musical form.

8.3 Decoder Strategies: Navigating the Manifold

The decoder within the broadcaster functions as the core of the interface’s co-creative logic. It operationalizes the discrete sampling findings from Chapter 7, transforming abstract topological movement through a high-dimensional hypersphere into audible, functionally coherent harmonic variations. To facilitate this active contextual navigation, the user can toggle between three primary sampling strategies. Each strategy represents a distinct theoretical and philosophical approach to exploring and defamiliarizing the possibility space of the J.S. Bach corpus.

By offering these three strategies, BroadcastJSB provides a tiered level of agency, with the option to either listen to the original chorale or create variations. In these variations, one is able to

⁷https://github.com/NadiaCarvalho/BroadcastJSB/blob/main/Create_Database.ipynb, last accessed January 29, 2026.

⁸<https://www.music21.org/>, last accessed January 29, 2026.

remain within the validated clusters of the corpus (NEAR), explore the geometric continuity of the latent model (MIDPOINT), or manipulate the functional momentum of the style itself (VECTOR). This implementation grounding ensures that the radio receiver metaphor is not merely a visual skin, but a functional representation of the deep topological organization explored throughout Part II of this thesis.

8.3.1 NEAR (k -Nearest Neighbors): The Shimmering Manifold

The NEAR mode implements the k -NN sampling strategy discussed in Section 7.2.2, looking for the k closest validated data points in the latent dictionary. The mapping between the tuning dial input (v , 0-100%) and the neighborhood size (k) follows a non-linear power curve defined by $k = v^3 \times 100$. This relationship models the topological density of the Bach manifold. At low values (10-50%), the system provides a narrow search radius that maintains the stability of the original signal with minimal variation. Musically, this preserves the functional core of the progression, such as cadences and harmonic rhythm, while introducing a dynamic “shimmering” re-harmonization where specific vertical sonorities are replaced by stylistically similar, common-tone chord substitutions. Conversely, as v increases, the rapid expansion of k reflects the loss of structural coherence characteristic of the manifold’s higher-entropy regions. In musical terms, this causes the strict parsimonious voice-leading and the original tonal center to dissolve, resulting in functionally ambiguous, highly chromatic sonorities that break away from Bach’s traditional harmonic grammar.

The introduction of drift through the tuning dial prompts the decoder to select neighboring coordinates within a dynamically expanding Euclidean search radius, jittering around the original Bach coordinate. Musically, this generates a dynamic re-harmonization. The system preserves the functional core of the progression, such as cadences and harmonic rhythm, but replaces specific vertical sonorities with stylistically related alternatives. By treating the latent space as a functional harmonic cluster, this mode maintains structural integrity through functional chord substitutions while continuously altering the surface-level vertical realizations.

8.3.2 MIDPOINT (Linear Interpolation): The Geometric Path

The MIDPOINT mode operationalizes the linear interpolation (LI) strategy. As established in Section 7.2.1, while LI can inadvertently navigate through “latent dead zones” over extended distances, it offers a distinct affordance for generating continuous, synthetic transitions between two structural anchors.

Operationally, the decoder bypasses the intermediate harmonic events of the original score, computing instead a straight-line vector between the user-defined initiation (A) and termination (C) points of a musical phrase. The vector is defined by the equation $z_t = (1 - t)z_A + tz_C$, which calculates a synthetic intermediate latent vector (z_t) at a specific weighted position (t) along the

continuous straight-line trajectory connecting a starting musical anchor (z_A) and an ending destination (z_C). Consequently, this navigation mode privileges geometric proximity and spatial smoothness over traditional functional harmonic logic. The resulting intermediate sonorities function as pure mathematical averages of the temporal anchors. Within BroadcastJSB, this mechanism generates synthetic transitions, harmonies that never appeared in the original 371 chorales but represent logically valid topological midpoints within the VAE’s learned grammar.

However, this reliance on static geometric pathing introduces specific performative constraints. Unlike other modes that dynamically expand their search radius based on the user’s tuning input, the MIDPOINT strategy is bound to the calculated vector, which can result in a phenomenon identified as a “geometric plateau.” Because the mathematical average between two distant tonal points frequently intersects the most statistically dense regions of the manifold, such as highly prevalent subdominant clusters, the interpolation often becomes trapped. Musically, this causes the system to repetitively sample the same centralized coordinates, occasionally neutralizing the necessary harmonic tension required for authentic cadences and pulling the sequence into functionally ambiguous zones. Thus, while the MIDPOINT mode demonstrates the mathematical continuity of the latent space, it highlights the inherent tension between raw spatial interpolation and the directional logic of common-practice voice-leading.

8.3.3 VECTOR (Angular Displacement): Tonal Momentum

The VECTOR mode operationalizes the k -NN+AD strategy, as developed in Section 7.2.3. Diverging from strictly proximity-based searches, this mode preserves the directional flow, or tonal momentum, of the musical sequence. Drawing from the geometrical alignment identified in Chapter 6, this strategy recognizes that the angular orientation of a vector in the latent space carries functional harmonic meaning (e.g., cadential movement toward the Dominant).

Algorithmically, the decoder filters the nearest neighbors to find the one that minimizes the angular deviation from the preceding voice-leading trajectory, calculating this alignment via cosine similarity. This allows for mathematically extrapolating the underlying syntactic logic of the Bach corpus. By following a specific angular vector, the sequence’s harmonic content is propelled into denser, highly chromatic regions of the manifold while rigorously preserving directional continuity. In this sense, it attempts to mimic the way a composer might extend a sequence or a modulation, using the manifold’s high-dimensional geometry to ensure that the drift feels intentional and structurally grounded rather than random.

8.4 Evaluation and Discussion

The evaluation of BroadcastJSB moves beyond objective reconstruction metrics, such as accuracy and KL-divergence (addressed in Chapter 5), to consider the subjective and performative efficacy of the topological interface. The discussion focuses on how the application, and its metaphor of the radio receiver, successfully bridges the black box of the VAE with the tactile requirements of musical practice, facilitating a unique mode of human-AI collaboration.

8.4.1 Navigation as Analysis

The discussion must start by addressing how the three decoder strategies (NEAR, MIDPOINT, VECTOR) act as tools for embodied music analysis. To demonstrate how the interface functions as an explainable chord substitution tool, we analyze the first phrase of Chorale *BWV 269* in G Major, depicted in Figure 8.5a.⁹ The phrase is constructed by 15 harmonic slices, following a standard common-practice trajectory, starting on the tonic *I* chord and landing on a dominant *V* (i.e., half cadence). This original score serves as our broadcast signal ground truth and establishes the rhythmic scaffold and functional grammar that the decoder strategies will subsequently detune. By adjusting the dial for the NEAR and VECTOR strategies, we can observe how the system preserves or dismantles the original Bachian signal.

Figures 8.6 and 8.7 illustrate the musical results of deviations at varying interference levels (e.g., dial at 0.1, 0.5, 0.9, and 1.0) using the NEAR and VECTOR strategies. Figure 8.5b illustrates the musical results of using the MIDPOINT strategy for the same chorale. As observed in Figure 8.6b (dial at 0.5), the system performs structural chord substitutions (e.g., replacing *iii*6 with *vi*4) that preserve the functional essence of Bach, proving that the latent space provides an explainable mapping of harmonic logic rather than random generative noise.

(a) Original.
(b) MIDPOINT.

Figure 8.5: Comparative harmonic analysis of the initial phrase of J.S. Bach’s Chorale *BWV 269* in G Major. (a) The original sequence, demonstrating a functionally varied common-practice progression. (b) The synthetic transition generated via the MIDPOINT (LI) strategy between the phrase’s temporal anchors. The generated trajectory illustrates a latent dead zone. Rather than navigating a functionally diverse path, the straight-line mathematical average becomes trapped within a statistically dense subdominant cluster, repeatedly sampling proximal coordinates (e.g., IV_{32}^5 and IV^7) and neutralizing the original cadential tension.

The NEAR strategy (k -NN) treats the latent space as a cloud of validated data points. At 10% tuning (Figure 8.6a), the system exhibits high structural fidelity, reproducing the initial functional progression identically ($I, I, IV6, vi, V6, I$). The first deviation occurs at the seventh chord, where the expected *iii*6 is replaced by *vi*4. This represents a stylistically coherent neighbor, preserving the underlying functional syntax while introducing localized harmonic variation. As the tuning input (v) increases, the power curve $k = v^3 \times 100$ rapidly expands the search radius. By 90% tuning (Figure 8.6c), the harmonic progression becomes significantly more complex (e.g., $ii\flat 65\flat 3$,

⁹All accompanying audiovisual examples demonstrating these strategies can be accessed at <https://nadiacarvalho.github.io/MutableMusicalInterface/#/software/broadcast-jsb>.

$IV6^b5$), yet it still utilizes chords that exist within the Bachian dictionary. At 100% tuning (Figure 8.6d), structural coherence completely dissolves. The final cadence is replaced by functionally ambiguous and highly chromatic aggregates such as $biii7_{42}$.

(a) NEAR $v = 0.1$. (b) NEAR $v = 0.5$.

(c) NEAR $v = 0.9$. (d) NEAR $v = 1.0$.

Figure 8.6: Harmonic variations of the initial phrase of J.S. Bach’s Chorale *BWV 269* in G Major generated via the NEAR (k -NN) sampling strategy. The subfigures illustrate the effect of a dynamically expanding Euclidean search radius driven by the non-linear tuning input (v). (a) At $v = 0.1$, the system exhibits high structural fidelity, preserving the underlying functional syntax with only minor localized variation (e.g., substituting iii^6 with vi^4). (b) At $v = 0.5$, a moderate spatial expansion introduces further stylistically coherent structural substitutions. (c) At $v = 0.9$, the progression grows substantially more complex, sampling dense, highly chromatic vertical sonorities (e.g., $ii\emptyset^{65b3}$ and $IV6^b5$) that strictly belong to the established corpus. (d) At $v = 1.0$, stylistic integrity completely breaks down; structural coherence dissolves, and the final cadential resolution is neutralized by functionally ambiguous, high-entropy aggregates such as $biii7_{42}$.

Unlike the proximity-based NEAR, the VECTOR strategy (k -NN+AD) prioritizes directionality. It interprets the user’s drift as a force applied to the musical trajectory. At 10% tuning (Figure 8.7a), the divergence is immediate. By the second chord, the sequence has already shifted from I to VI . Because this strategy respects angular displacement, it pushes the harmony onto a new path rather than jittering around the original. At 50% tuning (Figure 8.7b), we see the introduction of highly chromatic momentum chords such as $bII6\#53$ and $vii\emptyset6543$. These are not random, representing the system extrapolating the tonal direction of the previous voice-leading move into denser, higher-entropy regions of the manifold. Remarkably, across different high-value VECTOR trials, the system often converges on the same final sonority: $\sharp viob642$. Consequently, this demonstrates that the sustained tonal momentum ultimately forces the harmonic trajectory into a specific mathematical sink or functionally ambiguous interstation noise region, effectively exceeding the structural boundaries of traditional G Major.



Figure 8.7: Harmonic variations of the initial phrase of J.S. Bach’s Chorale *BWV 269* in G Major generated via the VECTOR (k -NN+AD) sampling strategy. The subfigures demonstrate the effect of preserving directional tonal momentum as the tuning input (ν) increases, driving the sequence along specific geometric trajectories rather than maintaining localized proximity. (a) At $\nu = 0.1$, the progression exhibits immediate structural divergence, establishing a new functional path (e.g., shifting from I to VI by the second chord). (b) At $\nu = 0.5$, the sustained angular displacement propels the harmony into denser, highly chromatic regions of the manifold, generating intentional momentum sonorities such as $bII_{53}^{6\sharp}$ and $vii\varnothing_{43}^{65}$. (c) At $\nu = 0.9$ and (d) $\nu = 1.0$, the continuous geometric extrapolation forces the sequence into a specific mathematical “sink” or functionally ambiguous “interstation noise” region. Across high-deviation thresholds, the trajectory reliably converges on the high-entropy aggregate $\sharp vi_{42}^{ob6}$, fully dissolving the traditional boundaries of the established diatonic grammar.

At last, the MIDPOINT strategy, utilizing LI to calculate a straight-line vector between a start point (z_A) and an end point (z_C), presents a first obvious drawback compared to the other strategies. It is not possible to use the dial to progressively detune the chorale. Because the point is always chosen next to the midpoint of the line between z_A and z_C , the results for the same sequence are always going to be very similar within various runs of BroadcastJSB. In our analysis of the G Major Chorale *BWV 269*’s first phrase, the MIDPOINT sequence exhibits a phenomenon called a latent dead zone. This is visually evident in the repetition of specific chords across the middle of the phrase: while the original Bach sequence moves through a varied progression ($V6$, I , $iii6$, V , vi , IV , $I752$, $IV42$, $vii\varnothing6$), the LI strategy becomes trapped in the subdominant space, oscillating primarily between $IV532$ and $IV7$. This repetition of $IV532$ (a specific vertical aggregate in the chorales dictionary) is not a malfunction but a reflection of the manifold’s topological density. In a high-dimensional space, the average of two points often lands in the most populated region of the learned grammar. Because subdominant-functioning chords (like IV) are statistically prevalent

and structurally central in the Bach corpus, the straight-line path between the phrase anchors frequently passes through this dense subdominant cluster. The result is a geometric plateau: the system selects the same validated coordinate multiple times because that coordinate is the nearest neighbor to the mathematical midpoint at several different steps along the vector.

8.4.2 The Defamiliarization of the Canon

A primary outcome of the BroadcastJSB interface is its capacity to defamiliarize the J.S. Bach corpus. By manipulating the tuning dial to introduce drift, the user no longer encounters the chorales as static historical artifacts, but rather as a fluid possibility space. Ultimately, the system facilitates a paradigm shift from passively listening to a score to actively exploring a style.

This defamiliarization is fundamentally driven by the user’s active contextual navigation through the available decoder strategies. Rather than relying on arbitrary algorithmic randomization, the system leverages the learned topological geometry of the manifold to stretch the boundaries of common-practice harmony. By deliberately toggling between the localized structural substitutions of the NEAR strategy, the synthetic geometric transitions of the MIDPOINT strategy, and the trajectory-extrapolating VECTOR strategy, the user actively dismantles the canonical rigidity of the original score. The interaction redefines the J.S. Bach corpus as a navigable terrain where the performer can intentionally push the harmony into denser, highly chromatic, or functionally ambiguous territories while remaining anchored to the system’s structural logic.

In traditional generative systems, the AI is often expected to produce a perfect imitation. However, in BroadcastJSB, the introduction of simulated spectral noise and analog artifacts in the audio engine (Section 8.2.2.3) serves a critical psychological purpose, signaling the structural integrity of the music, creating an aesthetic of imperfection. When the radio is perfectly tuned, the clarity of the additive synthesis mirrors the perfection of the counterpoint. As the signal becomes noisy through latent drift, the auditory friction mirrors the mathematical distance from the ground truth, creating a sensory feedback loop where we can hear the topology of the latent space.

8.4.3 Analog Radio as a Transparency Layer for Visualizing Topological Deviation

The choice to present a high-dimensional VAE through a vintage radio aesthetic is an exercise in interface transparency, within our established framework of contextual navigation. While many AI tools hide the topological distance between the input and the generated output, BroadcastJSB uses the signal strength bar, the latent proximity indicator, and the fragmenting ground-truth indicator to explicitly quantify this deviation.

The latent proximity indicator and the signal strength meter interfaces represent the algorithmic authority of the training data. When the indicator is stable and the proximity meter is at its peak, the user is interacting with unaltered corpus coordinates. When the indicator fluctuates, the system is visualizing its topological deviation. This visual feedback visually marks the boundary between the established stylistic canon and high-entropy latent regions. By explicitly rendering the distance from the source material, the interface ensures that we know precisely whether the

system is navigating the structural constraints of the training corpus or exploring highly divergent areas of the latent space.

At the same time, the JSB-seal, as a ground-truth indicator, functions as a vital layer of AI transparency. By exhibiting visual fragmentation during latent state transitions, the indicator explicitly visualizes the system’s topological deviation, effectively marking the boundary between the unaltered structural constraints of the training corpus and functionally ambiguous, high-entropy latent regions. This interactional transparency fosters true navigational agency, ensuring the user does not navigate blindly but can clearly perceive their mathematical distance from the source material.

These visual indicators create a transparent relationship between the user and the training data. Unlike opaque black box generators that obscure their topological distance from the source material, BroadcastJSB makes this spatial relationship explicit. This transparency enables a new form of navigational agency: the user knows precisely when they are operating within the structural constraints of the training corpus and when they are exploring highly divergent, functionally ambiguous zones of the latent space. Ultimately, this continuous visual feedback is a necessary condition for an exploratory dialogue, as meaningful contextual navigation can only emerge when the user is able to legibly engage with and decode the system’s contextually bounded possibilities.

8.4.4 Limitations and Creative Constraints

Although the k -d tree implementation (Section 8.2.2.2) successfully guarantees real-time performance, the evaluation simultaneously underscores the inherent limitations of a dictionary-based approach. By anchoring the system strictly to the discrete chords extracted from the 371 chorales, it compromises the fluid morphing capabilities characteristic of a fully continuous decoder. Consequently, rather than generating micro-variations of a single sonority, the algorithm is forced to transition discretely to the next available neighbor.

However, within the strict context of common-practice tonality, this inherent constraint operates as a necessary structural safeguard. It actively prevents the user from drifting into highly dissonant regions that would otherwise disrupt the stylistic illusion of a J.S. Bach chorale radio station. Thus, while this dictionary-based strategy successfully guarantees stylistic validity by restricting the generation of non-musical clusters, it simultaneously imposes a discrete, quantized quality upon the navigation. To resolve this, future iterations may investigate a hybrid architecture that facilitates continuous latent decoding, effectively bypassing the static dictionary, while strategically employing the k -d tree solely as a functional anchor to confine the generative drift within strict musicological boundaries.

8.5 Summary: BroadcastJSB as a Demonstration of Co-creative Affordance

This chapter has demonstrated the transition from theoretical latent space analysis to the creation of BroadcastJSB [Carvalho \(2026a\)](#). By implementing LI, k -NN, and k -NN+AD within a JavaScript environment, we have shown that real-time constraints can be met in the browser, providing a unique co-creative affordance that allows composers to interact with the tonal momentum of a corpus.

The development of BroadcastJSB concludes our exploration of Part II: Latent Spaces of Tonal and Structural Dimensions. We have moved from the representation of pitch (Chapter 5), through the mathematical alignment with phase spaces (Chapter 6), into the generative navigation of those spaces (Chapter 7), and finally to the implementation of a functional interface (Chapter 8).

Ultimately, the insights gathered here, demonstrating how latent spaces can be systematically mapped to physical gestures and mathematical frameworks, provide the foundational logic for Part III, where we extend these principles directly into the domain of timbre. Having successfully established the topological tonal interface through BroadcastJSB, the subsequent chapters will investigate how these generative manifolds can be expanded to encompass the multidimensionality of sound itself.

Synthesis to Part II:

Latent Spaces of Tonal and Structural Dimensions

The studies in Part II have examined tonal and structural aspects of music through the lens of latent space modeling. The exploration of harmonic progression representation and the strategies for sampling and aligning learned spaces with mathematical frameworks showcases how we can understand, navigate, and transform tonal musical ideas using computational environments. The experiments not only look at tonal latent spaces as valuable tools for analysis but also embrace them as interfaces for interaction. Thus, these reflections enable a transition from static representation to dynamic, co-creative practice.

At the foundation of this research, Chapter 5 explores the representation of tonal music within variational autoencoder (VAE) latent spaces. We examined how different symbolic encodings influence the model’s ability to uncover music-theoretical structures of J.S. Bach’s chorales. The findings reveal an important balance between reconstructive fidelity and cognitive alignment. Textual and categorical formats (i.e., ABC and MIDI) are great for retrieving information accurately, but spectral DFT encodings are better at reflecting how humans perceive tonal music. Ultimately, by prioritizing spectral periodicities, the resulting latent spaces serve as a dense, continuous map where musical similarity is encoded as spatial proximity. Thus, discrete symbolic notation is transformed into a navigable geometric manifold.

Chapter 6 then introduced Discrete Fourier Transform (DFT) phase spaces as the mathematical ground truth for tonal logic. By projecting pitch classes and pitch distributions into these spectral spaces, we uncovered the rotational symmetries that define functional harmony. A critical milestone was the empirical alignment found between MusicVAE’s latent space and the DFT spaces. This alignment confirms that the model’s unsupervised learning process naturally discovers structural principles consistent with established music theory.

Having established that the latent space is in tune with theory, Chapter 7 investigated how to navigate it in a meaningful manner. We contrasted discrete linear interpolation (LI), which often fails by crossing through musically incoherent regions, with data-driven strategies. k -nearest neighbors (k -NN) ensured stylistic validity by anchoring generation to real data points. In contrast, k -NN with angular displacement (k -NN+AD) preserved the tonal momentum and directional

logic of the manifold, bridging the gap between mathematical navigation and parsimonious voice-leading.

The practical culmination of these theoretical and experimental findings is realized in BroadcastJSB, whose implementation is detailed in Chapter 8. BroadcastJSB is a web-based latent space navigator that transforms the static chord dictionary into a performative instrument. By adopting a mid-century radio receiver metaphor, the interface provides a cognitive bridge for the user to interact with the high-dimensional latent space through the familiar gesture of tuning a radio. Chapter 8 discusses the design and development of BroadcastJSB. The instrument implements LI, k -NN, and k -NN+AD sampling strategies in real-time using a modular architecture that decouples high-speed spatial searches (via k -d trees) from additive audio synthesis (via Tone.js). As the user introduces drift from the original Bach chorale ground truth, the system provides instantaneous visual and auditory feedback, such as the flickering latent proximity indicator, that reflects the current mathematical distance from the corpus. In this environment, the latent space is no longer a black box but a transparent, mutable interface, allowing the composer to navigate the tension between historical canon and generative invention.

Ultimately, Part II demonstrates how computational representations of tonality can support both deeper analytical understanding and practical modes of creative engagement. By treating the latent space as a flexible, interactive canvas, we transition from the black box of deep learning to a transparent, guided compositional environment. This provides the necessary groundwork for the more embodied explorations of timbre, gesture, and motion that follow in Part III.

Part III

Timbre, Gesture, and Motion in Latent Spaces

Introduction to Part **III**:

Timbre, Gesture, and Motion in Latent Spaces

While Part **II** focused on the symbolic and tonal foundations of music through an experimental lens, Part **III** shifts the analytical and creative focus toward timbre, gesture, and musical motion. Because the concept of “gesture” is inherently broad, we must first define its scope: within this framework, gesture represents the embodied, expressive intent of the performer, including continuous, real-time fluctuations in breath pressure, articulation, and dynamic intensity that shape a musical phrase. If timbre is the acoustic color of the sound and motion is the structural trajectory of a phrase (such as parallel or contrary movements), gesture is the human driving force that physically navigates the machine’s latent space. Although these dimensions are vital for expressive performance and co-creative practice, they frequently defy traditional symbolic notation.

Given the evasive nature of timbre, often described simply as “everything beyond pitch and duration,” we study its encompassing dimensions through hands-on case studies involving professional musicians. The tenor saxophone serves as our primary vehicle for exploration due to its dense sonic palette and exceptional timbre versatility, which includes a well-documented repertoire of extended techniques like flutter-tonguing, slap notes, and multiphonics. The saxophone provides a rigorous, “high-stakes” test for the system. If a computational model can successfully navigate and synthesize its non-linear acoustic relationships, the framework can theoretically extend to any complex sound source. As a saxophonist, the author leverages over a decade of experience to coordinate these evaluations, providing a direct bridge between abstract neural activations and the embodied experience of the performer.

In Part **III**, we focus on four interrelated objectives:

1. **The Timbre Landscape and Latent Topography:**

Our first objective is to determine how Realtime Audio Variational AutoEncoder (RAVE) models organize complex instrumental timbres and whether these latent representations are musically explainable. In Chapter 9, we will analyze the tenor saxophone’s timbre landscape by embedding layered analyses of sonic characteristics into RAVE models to capture subtle nuances of tone and articulation. We will investigate the generality of this method by comparing findings across models trained on diverse datasets, aiming to reveal a structural hierarchy where initial latent dimensions capture broad attributes (like energy) while deeper dimensions represent specific instrumental features.

2. Strategies for High-Dimensional Navigation:

Building upon this structural understanding, our second objective is to transition from latent space analysis to the practicalities of interactive control. Chapter 10 explores methodologies for navigating these high-dimensional manifolds, proposing a system that integrates mixed music models. Moving beyond traditional navigation that relies solely on mathematical distances, we will investigate how sampling strategies within an AI-informed concatenative synthesis environment can manipulate timbre while preserving the morphological integrity of the source material.

3. Mapping Gestural Intent and Motion:

Transitioning from proximity-based sampling to intentional thematic development requires a formalized understanding of how musical ideas evolve. Therefore, our third objective is to link timbre qualities with musical motion and expressive intent. Chapter 11 introduces the Motiv dataset, a collection of expert saxophone recordings explicitly annotated to illustrate the geometric behavior of thematic elaborations. By formalizing a taxonomy of musical motions (parallel, oblique, and contrary), we will attempt to bridge the gap between neural activations and the structural logic of musical counterpoint, treating the latent space as a dynamic topography of gestures.

4. Design and Development of a Co-Creative System:

Our final objective is to operationalize these findings into a functional performance environment that overcomes the rigid constraints of fixed electroacoustic media. Chapter 12 details the development of Aethra, a co-creative system designed for real-time navigation. By integrating motion-aware sampling and latent weighting, Aethra facilitates a dynamic dialogue where the electronic component actively responds to the live musician.

Through these investigations, we aim to demonstrate that musical expression emerges from the complex interplay of timbre, physical gesture, and temporal motion. By modeling these elements, the chapters that follow will establish a flexible, personalized, and embodied performative ecosystem for the modern electroacoustic practitioner.

Chapter 9

The Timbre Landscape of the Tenor Saxophone

This chapter presents a comprehensive investigation into the explainability and creative affordances derived from navigating a specialized timbre latent space generated by Realtime Audio Variational AutoEncoder (RAVE) models. While timbre is central to musical expressivity, shaping both the identity of a performance and the interpretive choices of performers, its high-dimensional nature often eludes precise control in digital environments. To effectively manipulate an acoustic instrument’s complex sonic palette, we must understand its underlying topological connectivity. As demonstrated by [Caillon et al. \(2020\)](#), mapping these acoustic relationships into a continuous timbre latent space provides a foundation for exploring and interpolating novel sonic topographies in real time.

By analyzing the tenor saxophone through the lens of layered latent representations, this study attempts to decode the continuous acoustic relationships between tonal color, articulation, and expressive gestures. While these performative nuances are often extensively detailed in contemporary scores through prescriptive symbols and textual instructions, traditional notation inherently compartmentalizes them into discrete events ([Yee-King, 2022](#)). In contrast, this approach quantifies their underlying topological connectivity, translating detailed compositional instructions into a continuous, navigable mathematical manifold where the dynamic interplay of these elements can be precisely mapped and interpolated ([Vigliensoni et al., 2020](#); [Akten et al., 2020](#)).

We argue that the organizational logic of these neural networks provides a unique sonic topography. The findings illustrate how computational models can serve as exploratory tools, enabling composers and performers to map the latent space, experiment with hybrid timbre combinations, and discover novel expressive opportunities. This approach emphasizes that timbre is not a static property of sound, but a dynamic element that interacts fluidly with performance gestures and co-creative decision-making.

Through a combination of perceptual expert assessment and objective clustering analysis, this chapter seeks to bridge the gap between abstract neural activations and the embodied experience of the performer. By establishing a framework for timbre transparency, we lay the groundwork

for a co-creative interface where the machine’s learned representations are made accessible and musically intuitive for real-time improvisation. Crucially, this transparency provides the necessary methodological foundation to expand traditional fixed media into an interpretable, dynamic, and object-based system, effectively transforming the static tape into a reactive and performatively navigable environment.

The chapter is outlined as follows. Section 9.1 introduces the theoretical and technical framework of RAVE, detailing its capacity for continuous, high-fidelity timbre generation and latent space mapping. Section 9.2 delineates the experimental methodology, detailing the curated Sax-Timbre dataset alongside a dual evaluation approach that integrates perceptual expert assessments with objective mathematical clustering. Subsequently, Section 9.3 presents the findings of this twofold investigation, demonstrating how specific neural dimensions semantically organize and represent the intricate timbre nuances of the tenor saxophone. Finally, Section 9.4 synthesizes these observations, discussing their creative affordances and establishing the foundational logic required for transparent, real-time timbre manipulation in co-creative musical interfaces

9.1 The RAVE Model and Continuous Timbre Generation

The Realtime Audio Variational AutoEncoder (RAVE) is a powerful generative model designed for high-fidelity, real-time audio synthesis (Caillon and Esling, 2021). RAVE models leverage the capacity of variational autoencoders (VAEs) to learn audio data representations, enabling timbre transfer, dynamic manipulation, and synthesis in real-time, as well as the integration of AI-driven sound design into real-time production workflows (Nercessian, 2023). This integration is applied in various contexts, including live performances (Schroeder and Reuben, 2023; Armitage, 2023a; Privato and Muller, 2024) and interactive installations (Armitage, 2023b).

Unlike traditional models that rely on symbolic data or pre-computed descriptors, RAVE operates directly on the raw audio waveform to learn a highly efficient, compressed latent representation (z_{timbre}) (Caillon and Esling, 2021). Historically, raw waveform modeling has been computationally prohibitive due to the extreme temporal dimensionality of high-resolution audio, often forcing models to rely on low sampling rates or prohibitively slow autoregressive generation (Oord et al., 2016; Mehri et al., 2017). RAVE bypasses these limitations by leveraging a multiband decomposition of the raw waveform, which decreases the temporal dimensionality of the data and expands the temporal receptive field (Caillon and Esling, 2021). This architecture allows the model to generate 48kHz audio signals while operating up to 20 times faster than real-time on a standard CPU (Caillon and Esling, 2021).

To ensure both an organized latent space and high-quality audio output, RAVE utilizes a novel two-stage training procedure. The first stage focuses strictly on representation learning. During this phase, the model operates as a standard VAE, optimized using a multiscale spectral distance metric. Utilizing an amplitude spectrum-based distance ensures the model captures essential perceptual features without penalizing it for inaudible phase variations, which often hamper the learning process in raw waveform modeling. Once the representation is learned, the second stage

applies adversarial fine-tuning. Crucially, during this phase, the encoder is frozen. If the encoder were continually trained alongside the decoder during the adversarial stage, the delicate balance achieved by the evidence lower bound (ELBO) regularization would be disrupted, leading to a dramatic increase in the latent space’s dimensionality and destroying the compact representation established in the first stage. With the encoder safely frozen, the decoder is trained using a Generative Adversarial Network (GAN) objective, integrating a feature matching loss to significantly enhance the naturalness and fidelity of the synthesized audio (Caillon and Esling, 2021).

Critical for real-time interaction, RAVE incorporates a post-training analysis to balance reconstruction fidelity with representation compactness (Caillon and Esling, 2021). VAEs frequently suffer from latent collapse, where certain dimensions become uninformative. To resolve this, RAVE applies a Singular Value Decomposition (SVD) to isolate the informative dimensions from those that have collapsed to the prior distribution. This allows the system to strictly crop the latent space to its most essential dimensions without degrading the audio quality (Caillon and Esling, 2021). For co-creative music systems, this optimized compactness is vital, as it establishes a continuous, low-dimensional manifold where computational navigation is streamlined, and subtle gestural movements translate into perceptually smooth, dynamic transitions in instrumental timbre (Caillon and Esling, 2021).

9.2 Methodology: Layered Analysis of the RAVE Encoder

Despite their usefulness and real-time efficacy, RAVE models often remain opaque to users, making navigation of their complex, high-dimensional latent spaces challenging. Establishing transparent representations is essential for creative control, as it allows musicians to understand how a model interprets specific musical timbres and refine generative outputs accordingly (Chemla-Romeu-Santos and Esling, 2022).

In this sense, we propose to investigate a fundamental question: how do RAVE models internally organize complex instrumental timbres, and can these high-dimensional latent representations be rendered musically explainable? A particular focus of the experiment is directed toward a specialized timbre latent space meticulously designed to capture the intricate nuances of saxophone extended techniques. Within this context, timbre is defined as all auditory sensations excluding pitch, loudness, and perceived duration.

To pursue this inquiry, we employ a mixed-methods research design that iterates continuously across practice-based feedback and formal experimental evaluation. Specifically, we conduct a layered analysis of the RAVE’s encoder and decoder outputs utilizing a dual approach: a perceptual subjective assessment driven by expert musical feedback, alongside an objective mathematical examination using clustering metrics. This objective examination intends to quantify the extent to which the latent spaces recognize semantically meaningful timbre attributes. This is achieved by measuring intra- and inter-segment distances based on technique, register, dynamic, and their combinations across the dataset.

To ground this empirical investigation, we constructed the SaxTimbre dataset, which serves as the primary material for our evaluations. Performed by expert saxophonist Jorge Sousa and recorded within a controlled studio setting, this curated corpus comprises 191 individual audio samples yielding approximately ten minutes of recorded material. While generative models like RAVE typically require over three hours of audio to train from scratch, this compact corpus embraces a small data mindset. Rather than training a new model, this restricted 10-minute collection is deliberately designed to be projected onto pre-trained models, offering a practical way to evaluate how they internally organize these specialized techniques. Captured using a stereo pair of Sontronics STC-1S MT BK microphones, the dataset systematically organizes a broad spectrum of playing styles, ranging from conventional articulation to complex extended techniques, across four primary pitch registers and up to five distinct dynamic levels.

By integrating these methods and utilizing this specialized corpus, this investigation seeks to decode the structural logic of the tenor saxophone’s timbre landscape within RAVE, ultimately establishing a framework for timbre transparency that can be intuitively navigated in real-time co-creative interfaces.

9.2.1 Pre-trained Models and Training Data

To investigate the capacity for RAVE to generalize across diverse musical timbres, the experiment utilizes several pre-trained models from the Institute for Research and Coordination in Acoustic-s/Music (IRCAM) and the Intelligent Instruments Lab (IIL). These models, detailed in Table 9.1, differ in their underlying architectures, sampling rates, and the datasets upon which they were trained.

The RAVE models exhibit distinct structural characteristics that significantly influence their performance across various musical contexts. IRCAM’s models are based on the RAVE v2 architecture, which utilizes a continuous model employing the VAE objective function, specifically the evidence lower bound (ELBO) function, for regularization (Caillon and Esling, 2021). Conversely, the models from the Intelligent Instruments Lab (IIL) are trained on a modified iteration of RAVE v1, which is optimized for transfer learning and streaming inference within environments such as Max/MSP, Pure Data (PD), and SuperCollider (Intelligent Instruments Lab, 2023). These IIL configurations utilize causal convolutions and exclusively consist of encoder-decoder structures without priors. Technical specifications further differentiate the two groups, as IRCAM models operate at a sampling rate of 44.1kHz, while IIL models operate at 48kHz. Additionally, the IIL models typically feature a greater number of latent dimensions, averaging between 16 and 20, whereas IRCAM models range from 8 to 16 latent dimensions.

9.2.2 SaxTimbre: the Tenor Saxophone Timbre Dataset

As introduced previously, the primary goal of the SaxTimbre dataset curation was to capture the nuanced tonal intricacies inherent in the tenor saxophone’s sonic palette. The tenor saxophone is highly regarded for its timbre versatility, featuring a wide range of well-documented extended

Table 9.1: Pre-trained RAVE models employed in our experiments, including their architectural versions, sample rates, latent dimensionality, and training dataset descriptions.

Name	Team	RAVE	Rate	Dim.	Dataset	Ref.
Sol_full	IRCAM	v2	44.1kHz	8	Full Studio OnLine IRCAM database (more than 20,000 samples of extended instrumental techniques).	(Ballet et al., 1999; Cella et al., 2020)
Isis	IRCAM	v2	44.1kHz	8	ISiS Database of high-quality synthesized singing voices.	(Ardailon, 2017)
MusicNet	IRCAM	v2	44.1kHz	16	Musicnet database; 34 hours of chamber music across 11 instruments.	(Thickstun et al., 2017)
IILGuitar	IIL	v1	48kHz	16	Timbre-oriented dry recordings of electric guitar (plucking, scraping, etc.).	(Intelligent Instruments Lab, 2023)
SaxSoprano	IIL	v1	48kHz	20	Original soprano saxophone improvisation by Franziska Schroeder.	(Intelligent Instruments Lab, 2023)
VocalSet	IIL	v1	48kHz	16	VocalSet; high-quality recordings of 20 professional singers.	(Wilkins et al., 2018; Intelligent Instruments Lab, 2023)

techniques often used in contemporary Western Art Music (Kientzy, 1990; Weiss and Netti, 2010). The dataset was recorded by expert saxophonist Jorge Sousa in a controlled studio environment, using a stereo pair of Sontronics STC-1S MT BK microphones. To ensure compatibility with the chosen RAVE architectures, the audio, initially recorded at a 48 kHz sample rate, was downsampled to 44.1 kHz with a 24-bit depth.

The 191 curated sounds, totaling approximately ten minutes of recorded content, encompass a comprehensive spectrum of techniques, registers, and dynamics, which are summarized in Table 9.2. The dataset, deliberately designed to be projected onto pre-trained RAVE models, comprises conventional playing styles alongside unconventional extended techniques designed to expand the instrument’s sonic palette (Burtner, 2005). These include long notes performed with both non-vibrato and vibrato variations, as well as quarter-tone notes played with microtonal adjustments. Specific textural effects are captured through subtones, which feature an “airy and breathy” quality in the lower register, and flutter-tonguing, which produces a tremolo-like effect created by the tip of the tongue (Weiss and Netti, 2010). Percussive and complex timbres are represented by slap notes, characterized by sharp tonguing and bisbigliando, which is interpreted as a color trill (Weiss and Netti, 2010). Finally, the collection includes multiphonics, where two or more pitches sound simultaneously on the monophonic instrument, and growling, achieved through the simultaneous act of playing and singing.

Table 9.2: Collection of sounds encompassing the entire spectrum of the tenor saxophone (timbre space).

Technique	Register	Dynamic
Long non-vibrato notes	Altissimo	Very Loud
Long vibrato notes	High	Loud
Quarter-Tones	Medium	Medium
Flutter-tonguing	Low	Soft
Slap notes		Very Soft
Growling		
Bisbigliando		
Subtones		
Multiphonics		

The dataset recordings are distributed across four primary registers (Altissimo, High, Medium, and Low) to capture the unique spectral characteristics of each technique at different pitch levels. For techniques such as long non-vibrato, vibrato, and quarter-tone notes, we recorded two specific pitches per register: C7 and G6 (Altissimo), F6 and C6 (High), D5 and A4 (Medium), and C4 and B♭3 (Low). These notations are tailored for the tenor saxophone’s B♭ tuning. In concert pitch, these correspond to B♭5/F5, E♭5/B♭4, C4/G3, and B♭2/A♭2, respectively.

Specific constraints and characteristics influenced the sampling of extended techniques throughout the collection. Long subtone notes were exclusively applied to the low register of the instrument, specifically targeting the notes D4, C4, B3, and B♭3. Due to extreme execution difficulty

and infrequent usage in standard repertoire, flutter-tonguing and slap notes were omitted from the Altissimo register above G. Slap variations were instead incorporated across the High, Medium, and Low registers, featuring a distinction between a dry (*secco*) attack and a standard attack with clear pitch recognition. Bisbigliando position variations were established across all four registers, while growling long notes were gathered across the entire range, but without dynamic variations due to performance limitations in soft or very soft dynamics.

For multiphonics, the dataset documents five distinct sonic aggregates across different registers without dynamic variations, including 1) C $\sharp$ 5, D $\flat$ 6, and A6; 2) D $\sharp$ 5 and F $\sharp$ 5; 3) F $\sharp$ 4 and G5; 4) D $\flat$ 4, B $\flat$ 5, D $\flat$ 6, and F $\sharp$ 6; and 5) D $\flat$ 5, G $\flat$ 5, E6, and B6. Where execution permitted, each note was performed at five dynamic levels (Very Loud, Loud, Medium, Soft, and Very Soft) though efficiency constraints and physical limitations required that certain techniques in the High, Medium, and Low registers be studied without dynamic variation.

For further exploration, the full dataset is provided as supplementary material.¹ This dataset serves as a comprehensive resource for investigating the micro-timbre nuances of the tenor saxophone within deep learning frameworks (Carvalho and Sousa, 2024).

9.2.3 Evaluation Methodology

To identify a generative model capable of transparently organizing the complex multidimensionality of the tenor saxophone, and to explicitly decode which specific latent dimensions control discrete acoustic features (such as dynamics, pitch register, and extended techniques), a two-pronged evaluation strategy was adopted. Specifically, this experiment seeks to answer whether the inherently opaque latent space of pre-trained RAVE architectures can be deciphered into a mathematically structured and musically intelligible control surface for real-time improvisation. To achieve this, the framework combines professional musical assessment with objective clustering analysis to quantify exactly how individual neural network dimensions represent the instrument's timbre.

9.2.3.1 Perceptual Subjective Assessment

The first stage of the evaluation investigated the internal organization and creative transparency of the RAVE models through a systematic analysis of the latent space. By isolating and combining individual dimensions, we determined the specific musical and acoustic influence of each neuron on the reconstructed signal.

This was achieved through a meticulous examination of the encoder and decoder outputs to assess the transparency and sonic character of each internal representation. A professional saxophonist with over twenty years of experience conducted listening sessions for every sound within the dataset, providing an expert perspective on the preservation of instrumental identity. To facilitate this, the output of each neuron in every latent dimension was converted into an audio signal

¹<https://doi.org/10.7910/DVN/WSFBIJ>, last accessed January 29, 2026.

and evaluated individually. We integrated the `nn_tilde` object² within the PD programming environment,³ allowing for real-time monitoring and manipulation of latent activations as they were decoded back into the acoustic domain.

During these sessions, the recorded saxophone sounds were passed through the encoder of each model, where the resulting latent representations were systematically manipulated before decoding. This process followed two primary methods: individual dimension isolation and combinatorial analysis. In the isolation phase, all latent dimensions were muted except for one specific dimension (x), forcing the decoder to reconstruct the sound based solely on the information captured by that single neuron. This enabled the identification of specialized dimensions, governing specific attributes such as breath noise, fundamental pitch, or the percussive attack of a slap note. Subsequently, we explored combinations of latent dimensions to observe their interactions and determine whether certain timbre features were represented globally or localized to specific clusters of neurons.

This analysis aimed to identify the sensitivity of specific latent dimensions to critical sonic attributes, including the sharpness of articulation, the stability of pitch modification, and the presence of spectral noise. By systematically listening to these reconstructions, the evaluator could pinpoint exactly where the model began to generalize features or lose the specific nuances of extended techniques, such as multiphonics or subtones. This perceptual mapping was essential for determining the reactivity of each model, defining how closely the internal architecture mirrored the physical and expressive reality of the instrument. To systematically categorize this reactivity into the qualitative labels (High, Moderate, Low, and Very Low) utilized in our results, the expert evaluation relied on three specific criteria:

- **Dimensional Activation:** The ratio of latent dimensions that actively respond to the audio input versus those that remain completely inert (indicating domain mismatch or latent collapse).
- **Acoustic Feature Sensitivity:** The model's capacity to discern and dynamically reproduce micro-variations in articulation, pitch stability, and spectral noise.
- **Register and Technique Fidelity:** The model's ability to maintain the instrument's timbre identity across all pitch registers without exhibiting latent dead zones, particularly during the execution of complex extended techniques such as multiphonics or subtones.

Ultimately, this qualitative ground truth provided the necessary context to interpret the subsequent mathematical clustering data, bridging the gap between high-dimensional latent points and the embodied musical experience.

²https://github.com/acids-ircam/nn_tilde, last accessed January 29, 2026.

³<https://msp.ucsd.edu/software.html>, last accessed January 29, 2026.

9.2.3.2 Objective Analysis of Timbre Clusters

The second phase of our evaluation focuses on quantifying the organizational logic of the RAVE latent spaces through mathematical clustering metrics. The goal of this objective examination is to determine if the models' internal representations align with semantically meaningful musical categories (i.e., technique, register, and dynamics), thereby validating the observations made during the perceptual subjective assessment.

To perform this analysis, the latent representations of the audio samples were partitioned into segments of approximately 40 milliseconds. These segments were then grouped into clusters based on four primary categorical frameworks: 1) Technique (e.g., slap, flutter-tongue, long note); 2) Register (Altissimo, High, Medium, Low); 3) Dynamic (from Very Soft to Very Loud); and 4) Combinations of the above attributes.

To evaluate the separation and compactness of these clusters, two metrics were employed: 1) Davies-Bouldin score (DB), and 2) Dunn index. The DB score calculates the average similarity between each cluster and its most similar one (Davies and Bouldin, 1979). Similarity is defined as a ratio of within-cluster dispersion to between-cluster separation. In this context, a lower DB score indicates better-defined clusters that are both compact and well-separated, suggesting that the model has successfully learned to distinguish between different musical attributes. The Dunn index is the ratio of the minimum inter-cluster distance to the maximum intra-cluster diameter (Dunn†, 1974). It serves as a measure of cluster density and isolation. Unlike the DB score, a higher Dunn Index indicates superior clustering, where the groups are tightly packed and significantly isolated from one another.

A critical component of this methodology involved comparing the clustering results between normalized and non-normalized audio data. By removing the influence of absolute volume through normalization, we were able to determine whether a model's ability to cluster sounds was driven purely by energy (loudness) or if it was truly recognizing the spectral fingerprint of the saxophone's timbre.

This mathematical framework allows us to pinpoint exactly which latent dimensions in models like MusicNet or VocalSet are most efficient at organizing specific attributes. For instance, if a specific latent dimension shows a significantly lower DB score for the technique category than other latent dimensions, it identifies that latent dimension as the most semantically transparent for controlling extended techniques in a real-time performance interface.

The implementation code for the latent dimension-by-dimension evaluation and clustering analysis has been made available in a dedicated repository.⁴ This includes the scripts used for quantifying intra- and inter-segment distances, as well as the perceptual assessment framework.

⁴<https://github.com/NadiaCarvalho/SMC-TimbreLandscape>, last accessed January 29, 2026.

9.3 Results and Discussion

The evaluation of the RAVE models reveals critical insights into how neural networks represent the intricate timbre nuances of the tenor saxophone. The findings are presented in two parts: perceptual observations from expert listening and objective data derived from clustering metrics.

9.3.1 Perceptual Subjective Assessment Results

The subjective evaluation, conducted through systematic dimension-toggling and granular listening sessions, provided a detailed view of how specific RAVE architectures and training corpora influence the perception of tenor saxophone timbres. The MusicNet and VocalSet models emerged as the most responsive, demonstrating a high degree of timbre intelligence by discerning micro-variations across the instrument’s entire register. Conversely, the SaxSoprano model proved remarkably ineffective. Despite its woodwind-based training data, 11 of its 20 latent dimensions remained inert when processing the tenor input.

This 55% inactivity rate suggests a combination of latent collapse and domain mismatch. In the context of VAEs, latent collapse occurs when the KL divergence term dominates the loss function during training, leading the model to ignore specific dimensions in favor of a lower-complexity representation. This was likely exacerbated by the spectral divergence between the two instruments: the tenor’s deep fundamental frequencies and complex sub-harmonics represent an “out-of-distribution” signal relative to the high-frequency bias of a soprano-only dataset. Essentially, the encoder likely discarded these unfamiliar tenor features as noise. In contrast, the more robust models (MusicNet and VocalSet) consistently localized high-level features (e.g., amplitude envelopes and percussive slap articulations) within the first latent dimension, suggesting a more hierarchical and semantically accessible organization of sound.

In our examination of the VocalSet model comprising 16 latent dimensions, we found it to be a balanced model, with latent dimensions responding to sound samples across all registers (from Low to Altissimo). A notable exception was latent dimension 16, which showed greater reactivity specifically to the Low register, a phenomenon captured in the spectral analysis shown in Fig. 9.1. While several latent dimensions within this model reacted to multiphonics, as illustrated in Fig. 9.2, the second and third multiphonics in our dataset failed to elicit any response. Latent dimension 1 emerged as the most reactive overall, while latent dimension 11 displayed high reactivity to the Altissimo register, and latent dimension 13 responded solely to the note C across various registers.

In the MusicNet model specifically, latent dimension one excelled at capturing distinctive attacks, while latent dimension five demonstrated a unique capacity to modify the pitch of the reconstructed sounds. Other models showed more localized functionality, such as the Isis model, which displayed heightened responsiveness to the lower register notes $B\flat_3$, B_3 , and C_4 , as well as louder dynamic levels. Conversely, the Sol_full model exhibited very little reactivity, with six of its eight latent dimensions failing to show any discernible response to the sound samples.

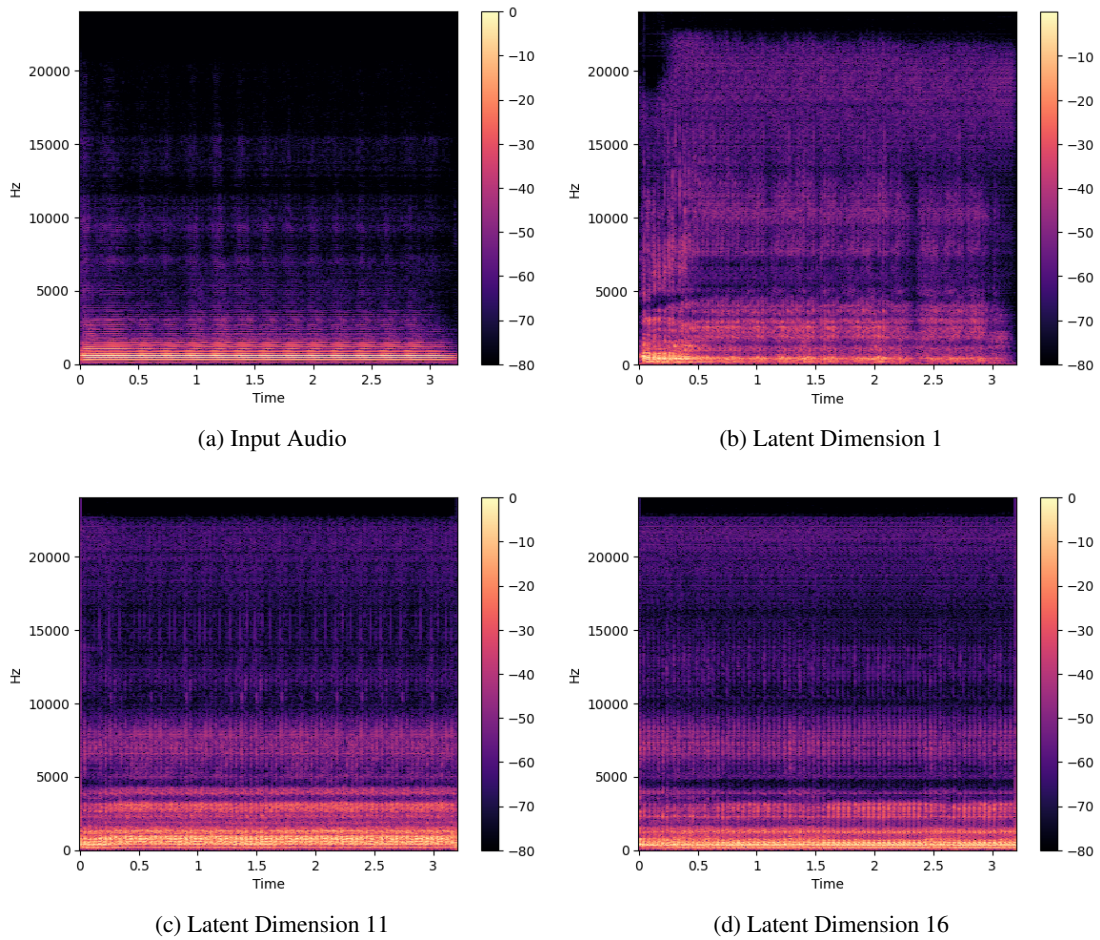


Figure 9.1: Comparison of the input spectrogram for a soft low C4 recording against the reconstructed outputs from latent dimensions 1, 11, and 16 of the VocalSet latent space. The visualization highlights the progressive transformation of spectral content and the specific low-register reactivity observed in layer 16.

Despite the general reactivity found in the top-performing models, several limitations were observed during the listening tests. Most models proved ineffective at responding to or distinguishing between various multiphonics, and even the highly reactive VocalSet model failed to elicit a response for certain complex multiphonic samples. Furthermore, many deeper latent dimensions in models like IILGuitar Timbre and MusicNet appeared to be ineffective, failing to discern most sounds or react to stimuli altogether. These perceptual insights suggest that while the initial latent dimensions of the RAVE encoder capture broad musical attributes, such as energy and attack, the deeper latent dimensions often represent highly specialized features that may not align with the semantic expectations of a performer navigating the latent space.

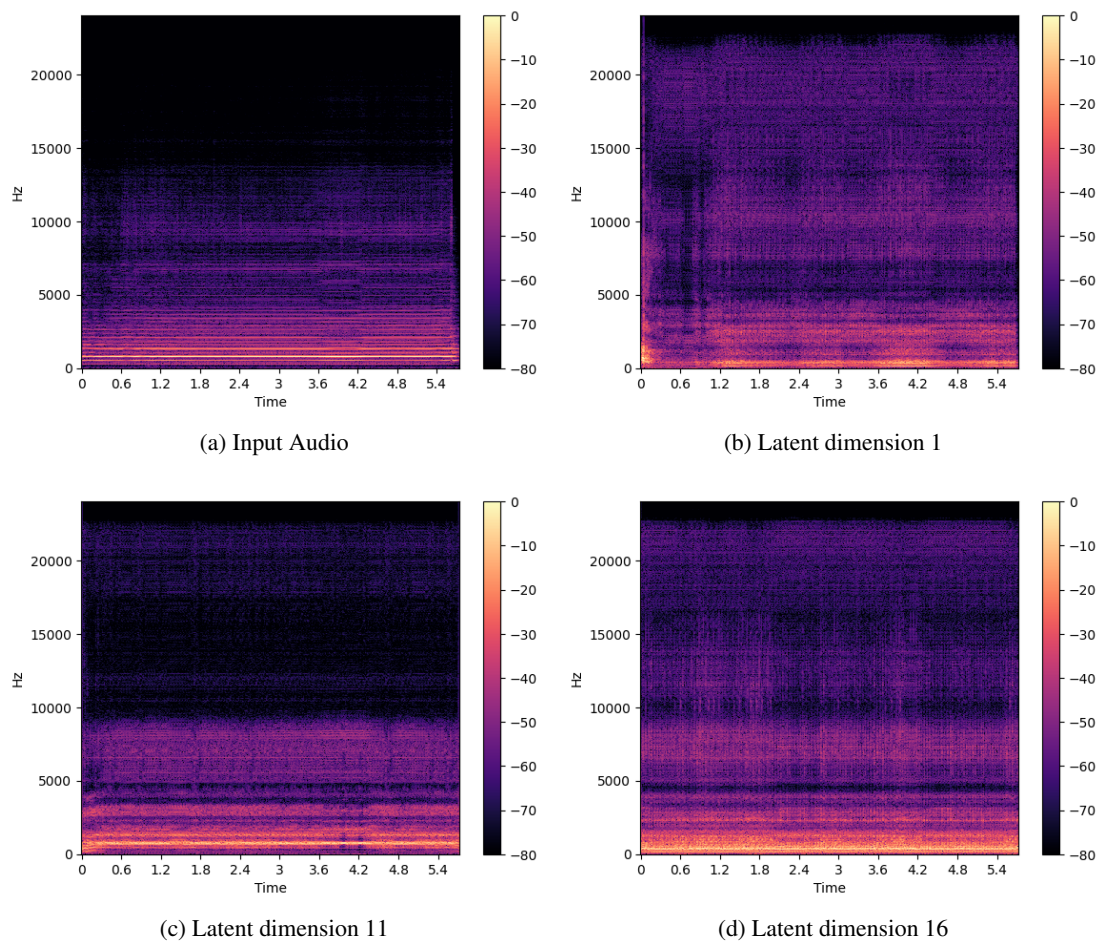


Figure 9.2: Comparative spectrograms of the input audio for multiphonic aggregate no. 5 alongside reconstructions decoded from latent dimensions 1, 11, and 16 of the VocalSet latent space. The visualization demonstrates the varying degrees of spectral fidelity across the model hierarchy and the specific challenges of representing non-harmonic partials in deeper encoder latent dimensions.

9.3.2 Results of Objective Clustering Analysis

The objective evaluation of the latent spaces provided a quantitative map of how timbre is structured within the RAVE models. By analyzing the Davies-Bouldin (DB) and Dunn indices across various architectures, we observed clear trends in how neural networks prioritize musical information. These findings, summarized in Table 9.3, provide a mathematical foundation for the perceptual observations noted previously.

Table 9.3: Comparative analysis of clustering performance across all RAVE model latent dimensions. Optimal values for each categorical framework and metric are indicated in bold. The labels N and NN distinguish between evaluations performed with normalized and non-normalized audio input, respectively, highlighting the impact of signal amplitude on latent organization.

Model	Metric	Technique		Register		Dynamic	All
		N	NN	N	NN		
Sol_full	DB	9.3978	63.1686	3.1654	71.0053	61.2390	23.3833
	Dunn	0.0001	0.0003	0.0002	0.0003	0.0002	0.0014
Isis	DB	7.8626	36.9219	2.5262	17.9766	30.7508	18.4759
	Dunn	0.00002	0.0003	0.0001	0.0005	0.0001	0.0006
MusicNet	DB	6.9871	27.2322	3.4207	22.0426	35.3139	14.3542
	Dunn	0.00002	0.0003	0.0001	0.0005	0.0003	0.0016
IILGuitarTimbre	DB	46.4832	69.0797	25.9224	40.3061	58.2117	35.3975
	Dunn	0.0002	0.0002	0.0005	0.0003	0.0001	0.0003
SaxSoprano	DB	8.8365	23.6323	3.1240	5.8060	14.4397	9.2430
	Dunn	0.0003	0.0004	0.0004	0.0005	0.0002	0.0034
VocalSet	DB	11.4939	12.7429	2.5283	2.9858	8.4275	5.5918
	Dunn	0.0003	0.0003	0.0002	0.0005	0.0003	0.0024

The results in Table 9.3 confirm that the highest degree of semantic organization occurs in the early stages of the encoder. Specifically, the DB scores for dynamic levels were at their lowest in the initial latent dimensions, indicating that the network immediately isolates signal energy as a primary organizational feature. This mathematical result aligns with the perceptual assessment, where the expert listener identified latent dimension one as the most reactive to volume and attack.

Figure 9.3 showcases the outcomes of the cluster analysis, organized latent dimension by latent dimension. Due to the close resemblance of Dunn index results across all metrics (approximately 1×10^{-5}), only Davies-Bouldin score results are presented in the figure. Several discernible patterns emerge from these results, complementing the findings of the listening test. Specifically, dynamics are consistently better generalized in the initial latent dimension of the model, except in the case of the IILGuitarTimbre and VocalSet models, where they are observed in the second and sixth latent dimensions, respectively.

Combined sounds are not adequately clustered at any model latent dimension, likely due to the limitation of a single latent dimension in capturing the diverse timbre attributes of entire sounds.

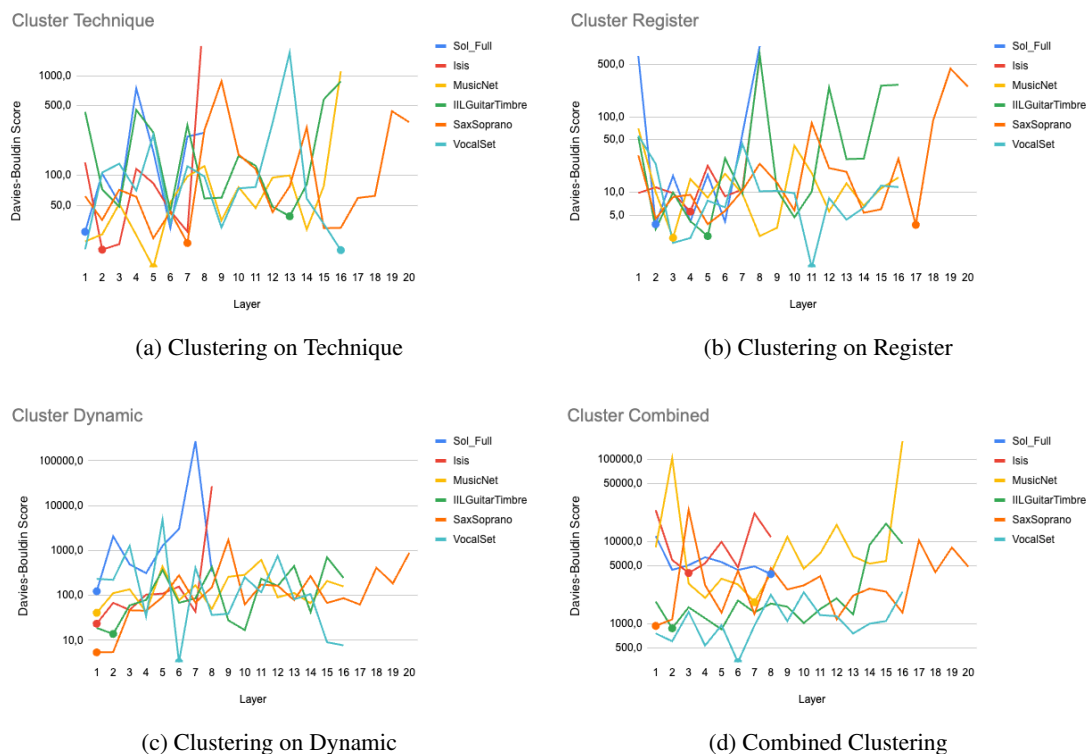


Figure 9.3: Davies-Bouldin Score Results by latent dimension. The best results are marked with a circle. All audio is normalized for the cluster categories corresponding to extended techniques and pitch registers.

The pitch register is generally well perceived across the models, although not consistently at the same latent dimension in every model. At the level of extended techniques, MusicNet, closely followed by VocalSet and Isis, are the best models at semantically perceiving timbric changes within a single latent dimension.

As shown in the comparative data in Table 9.3, the MusicNet model achieved the most compact clusters for specialized techniques when audio was normalized, yielding the best DB score of 7.01. However, its lower Dunn Index suggested that these clusters, while internally consistent, were positioned relatively close to one another in the latent space. In contrast, the SaxSoprano model demonstrated a high degree of mathematical organization despite its lower perceptual reactivity. When audio was normalized, it achieved the highest overall cluster isolation with a Dunn Index of 0.0034, as indicated in Table 9.3, revealing very well-isolated clusters for different playing techniques. This suggests that while the model may not sound as reactive to the expert ear due to its training on a different instrument, its latent space is structurally robust and could potentially be remapped for creative use.

The comparison between normalized and non-normalized data proved essential for understanding the models' timbre-transparency. For all models, normalization significantly improved the DB scores for both technique and register clustering. Without normalization, the models tended to cluster sounds primarily by their loudness (e.g., grouping a loud long note with a loud slap note),

effectively masking the more subtle spectral differences of the extended techniques. These findings underscore the importance of pre-processing for performers who wish to utilize latent spaces for navigating complex timbre landscapes, rather than relying on simple volume-based interactions.

9.4 Summary: Creative Affordances and Real-Time Manipulation

The comprehensive examination of RAVE models in this chapter provides crucial insights into how neural network-based architectures represent the complex timbre attributes of the tenor saxophone. Through the dual evaluation of subjective listening and objective clustering, several key findings have been established regarding the transparency and functionality of these latent spaces. As summarized in Table 9.4, among the various models studied, MusicNet and VocalSet demonstrated the highest reactivity to tenor saxophone samples, particularly within the initial latent dimensions of the network. The first latent dimension consistently emerged as the most robust for representing changes in dynamics, while sensitivities to pitch and register variations were distributed inconsistently across different latent dimensions and models. Furthermore, the study highlighted that the absence of audio normalization significantly hinders the models' ability to effectively distinguish between extended techniques and pitch registers.

The insights gleaned from this timbre landscape offer a clear path forward for enhancing control and transparency in co-creative music frameworks (Carvalho et al., 2024). By identifying the latent dimensions that govern specific musical attributes, such as dynamics or percussive attacks, interface mapping strategies can be developed that enable performers to interact with these parameters more intuitively. The superior reactivity of the MusicNet and VocalSet models suggests that datasets featuring high timbre diversity provide a more flexible latent space for processing specialized instrumental techniques than models trained on narrower datasets. Furthermore, these findings serve as the foundation for developing new software tools, where integrating the SaxTimbre dataset with motion-based datasets fosters a collaborative environment in which the energy of a physical gesture directly shapes the evolution of synthesized saxophone timbres.

The ability to successfully isolate and decode the impact of specific encoder latent dimensions holds profound implications for generative musical practice, particularly in real-time co-creative performance. By observing which latent dimension is responsible for a specific feature, a composer or performer gains the ability to manipulate that feature independently in the latent space, interacting with musically intuitive controls derived from the learned structure rather than abstract synthesis parameters. The continuous nature of the latent space facilitates seamless morphing between distinct saxophone timbres, such as a clean tone and a growling tone, while navigating the periphery of the learned manifold enables controlled extrapolation into novel sounds that retain the characteristic texture of the instrument. This layered analysis provides a degree of explainability for the RAVE model, showing precisely which musical information the model prioritizes at each stage of compression, a significant step toward developing transparent and trustworthy AI tools.

Building upon this structural understanding of the timbre landscape, the following chapter transitions from latent space analysis to the practicalities of interactive control. As we move

Table 9.4: Qualitative and quantitative summary of RAVE model reactivity to tenor saxophone timbre. The table synthesizes expert perceptual assessments with mathematical clustering metrics (Davies-Bouldin and Dunn Indices) to define the creative transparency and structural limitations of each architecture.

Model	Reactivity	Key Strengths	Observed Limitations
MusicNet	High	Achieved the highest semantic accuracy for technique (DB: 6.9871). Layer 1 excels at attack detection, while Layer 5 provides unique pitch-shifting affordances.	Exhibits high cluster overlap, making it difficult to isolate specific techniques without accidental morphing.
VocalSet	High	Most balanced performance across registers. Layer 16 is uniquely optimized for low-register microtimbres like subtones.	Demonstrates spectral gaps where specific complex multiphonic aggregates (No. 2 and 3) fail to trigger any latent activation.
Isis	Moderate	Superior cluster separation for register-specific notes (<i>B♭3</i> to <i>C4</i>). Provides high transparency for dynamic-heavy improvisational play.	Significant latent dead zones in the Altissimo range; effectively loses tracking of instrumental identity above F6.
IILGuitar	Moderate	High fidelity in capturing the noise component of <i>secco</i> slap notes. Specialized layers allow for isolation of percussive transients.	Many deeper dimensions are functionally redundant or silent, limiting the total useful navigation area of the manifold.
Sol_full	Low	Preserves high-frequency spectral detail due to its 44.1kHz architecture.	High degree of latent collapse; 75% of the dimensions (6 of 8) do not respond to saxophone stimuli.
SaxSoprano	Very Low	Structurally robust clusters with the highest mathematical isolation (Dunn Index: 0.0034).	Extreme domain mismatch leads to an 11/20 inactivity rate (55%); non-transparent for real-time instrumental play.

forward, we explore both traditional and experimental methodologies for navigating these high-dimensional spaces in performance contexts. By examining how different mapping strategies and interaction paradigms influence the exploration of these learned manifolds, we aim to bridge the gap between static neural representations and the fluid, gestural nature of live musical improvisation.

Chapter 10

Navigating High-Dimensional Timbre Latent Spaces

Following the structural analysis of various pre-trained RAVE models’ latent dimensions in Chapter 9, this chapter explores the practical methodologies for navigating these high-dimensional manifolds. This research transitions from latent space theory to implementation, specifically investigating how sampling strategies can be leveraged to manipulate timbre while preserving the morphological integrity of the source material. In this context, the study moves beyond traditional navigation, which often relies solely on perceptual distances between sound segments, to propose a system that integrates mixed music models, empowering musicians to make more stylistic and performative choices.

By viewing the latent space as an AI-informed corpus-based concatenative synthesis (CBCS) environment, we investigate how computational agents can model divergences from established musical norms. This approach is designed to enhance rather than supplant the performer’s creative agency, providing a structured foundation for expressive exploration. The focus centers on Jesús Villa-Rojo’s *Lamento* for tenor saxophone and tape as a primary case study (Villa-Rojo, 1989). This work is particularly suitable due to its temporal freedom and intricate timbre nuances, which offer a rich ground for testing navigation mechanisms within the latent domain.

Aiming to develop a co-creative system, we leverage RAVE models to treat timbre as a primary navigable dimension. By implementing sampling strategies that account for both spatial proximity and temporal trajectories, we aim to introduce interpretable structures into these high-dimensional spaces. This evolution toward motion-aware navigation supports dynamic composition and expressive real-time interaction, allowing the machine’s learned representations to become an intuitive extension of the performer’s musical vocabulary.

The remainder of this chapter is structured as follows. Section 10.1 examines the inherent difficulties of real-time timbre navigation within high-dimensional generative models, particularly addressing morphological drift and the constraints of continuous latent control. To mitigate these issues, Section 10.2 introduces three algorithmic sampling strategies for structured exploration:

k -nearest neighbors (k -NN) binning, distance-based (BD) binning, and distance with aligned trajectory (BD-AT). Section 10.3 details the experimental methodology, utilizing Jesús Villa-Rojó’s *Lamento* as a practical case study to evaluate these strategies across varying divergence parameters through both perceptual assessments and objective similarity metrics. Subsequently, Section 10.4 presents the findings of this comparative analysis, illustrating how the different navigational logics impact temporal continuity and feature preservation during live generation. Finally, Section 10.5 synthesizes these observations, defining the threshold of morphological coherence and demonstrating the necessity of trajectory-aware navigation for robust co-creative systems.

10.1 The Challenge of Timbre Navigation in Real-Time Systems

While the Realtime Audio Variational AutoEncoder (RAVE) provides a high-fidelity, continuous latent space for timbre (z_{timbre}), navigating this manifold in a musically meaningful manner during live performance presents significant challenges. In the context of electroacoustic mixed music, the primary obstacle is often the restriction of the performer’s freedom and intentionality. Interactive music systems, such as the free improvisation models proposed by Winkler, attempt to reduce these limitations. However, implementing such systems, particularly those aiming for the co-creativity found in autonomous agents like Somax2 (Borg et al., 2022; Fiorini and Malt, 2023), remains one of the primary obstacles in machine-generated music due to the difficulty of aligning stylistic coherence with real-time interpretive choices.

A fundamental challenge in this study arises from the constraints of the training data. A RAVE model typically requires a substantial dataset, often exceeding three hours of audio, to generate a sufficiently significant and high-quality output. Due to a lack of such extensive audio material for specific, specialized corpora, as is common in Western Art electroacoustic music, where each work is typically at most half an hour and characterized by a particular language of extended techniques, we utilize a strategy where a restricted corpus is projected onto a generic RAVE model that was not trained on the corpus itself.

Navigating this latent space involves moving between discrete regions that correspond to the specific musical samples of a curated work. This process is conceptually analogous to CBCS, where audio units are selected based on their “proximity to a target position in the descriptor space” (Schwarz et al., 2007). In traditional CBCS, this descriptor space is strictly feature-based, meaning human developers must manually curate and prioritize specific acoustic parameters, a significant bottleneck that often limits the system’s flexibility. Deep generative models resolve this curation challenge because their neural networks autonomously establish complex, non-linear relationships directly from the raw audio to form the latent space. However, projecting a specialized instrumental corpus onto a generic, pre-trained neural model introduces a new challenge: the risk of morphological deviation. Because the model’s latent representation was optimized for the broad statistical distribution of its training data rather than the specific nuances of a particular musical work, synthesized outputs may gradually deviate from the original sonority while still maintaining a relationship to the source material’s initial morphology.

In the broader context of high-dimensional audio synthesis, the relationship between latent space proximity and musical coherence presents a significant non-linear challenge. As navigation parameters are adjusted to encourage greater variation, effectively moving further from a specific target or reference point, sampling algorithms frequently encounter this drift. This phenomenon occurs when the navigation trajectory enters statistically dense regions of the manifold that lack the semantic or morphological structure of the original input. Rather than producing meaningful musical development, this drift manifests primarily as a degradation of signal integrity and timbral coherence, resulting in unwanted acoustic artifacts, the loss of the instrument’s distinct spectral signature, and an eventual dissolution into unstructured noise.

Furthermore, the high dimensionality of the RAVE latent space, typically comprising more than eight dimensions, as seen in the MusicNet model, further complicates precise manual navigation. Adjusting weights for each latent dimension in real-time is a cognitively demanding task that can detract from the musical performance. Consequently, there is an inherent trade-off between the complexity of the sound manipulation and the immediacy required for live interaction. The challenge lies in developing intuitive control mechanisms that can handle this underlying high-dimensional complexity while providing the performer with a structured, musically coherent foundation for exploration.

10.2 Algorithmic Navigation Strategies for Latent Space Exploration

To address the challenges of morphological drift and the lack of contextual awareness, three distinct sampling strategies were developed to prioritize the structural integrity of the original timbre while allowing for controlled divergence.

Before detailing these strategies, it is critical to explicitly define the temporal dimension and the fundamental audio units utilized in this context. Unlike the symbolic MusicVAE architectures employed for harmonic navigation in Part II, this experiment utilizes a pre-trained RAVE model (specifically the one trained on the VocalSet dataset), which operates directly on raw audio waveforms at a 48 kHz sample rate. The model compresses this raw audio into 16-dimensional latent vectors. In this context, an individual “audio unit” refers to a single, discrete latent frame, corresponding to a temporal segmentation of approximately 43 milliseconds (or 23 frames per second of audio).

Because RAVE generates a mathematically continuous and extremely dense manifold, searching for optimal audio units frame-by-frame during a live performance introduces a severe computational bottleneck. To resolve this and ensure real-time efficiency, the following algorithmic strategies organize these continuous latent units into discrete categories (or bins). By discretizing the continuous space into these manageable categories, governed by a primary closeness parameter, C , the system circumvents the need to execute exhaustive distance calculations across the entire manifold for every 43-millisecond unit. This categorical reduction provides the structured, low-latency foundation required for expressive real-time exploration.

This parameter C serves as a control input, ranging from 0 to 1, that determines the degree of proximity to the original musical work. At a value of $C = 0$, the system maintains the integrity of the original tape, ensuring the output is identical to the source material. As the value of C increases toward 1, the navigation strategies are permitted to sample audio units that deviate significantly from the reference sequence.

This parameter operates as the gateway for performance freedom in the development of a co-creative system, which will be detailed later in Chapter 12. By adjusting C , the performer can transition from a state of perfect synchronization with the pre-recorded material to a state of high divergence, where the system generates dissonant, rhythmic, or timbre variations. The efficacy of C is intrinsically tied to the chosen strategy, as it dictates which category of latent units the algorithm targets during the sampling process.

10.2.1 k -Nearest Neighbors (k -NN) Binning

The k -nearest neighbors strategy implemented for timbre navigation is closely related to methods used in tonal music latent spaces, as discussed in Chapter 7. The k -NN algorithm serves as a versatile approach for both classification and regression within supervised learning, effectively adapting to diverse datasets by determining a unit's value based on the average of its k nearest neighbors in the feature space.

In this context, the strategy focuses on capturing distinct timbre attributes by evaluating their spatial proximity in the embedding landscape. To extract a specific audio unit, categories are created using the k -NN algorithm to pinpoint the k closest neighbors within a meticulously defined embedding space. The value of k is determined by dividing the total number of units in the space by the desired number of categories. If the total units are not a multiple of the categories, the closest categories retain more units.

This method, illustrated in Figure 10.1, entails measuring distances (e.g., cosine or Euclidean distances) between a chosen unit and every other frame in the embedding neighborhood to isolate the k neighbors assigned to each category.

Within each category, a selection process utilizes a random weighted distribution method where weights correspond to the distance from the original musical object. For instance, in an environment with 25 categories, the value of C increases in increments of 0.04. If k is approximately 25 and $C = 0.2$, the selected unit will belong to the fifth bin. Potential units are weighted based on their proximity to the selected unit, from least to most distant, before a final random selection is made. Importantly, bin zero is uniquely designated to always lead back to the original object, ensuring that setting $C = 0$ perfectly maintains the integrity of the original musical work.

10.2.2 Distance-Based (BD) Binning

While k -NN binning provides a structured way to categorize the latent space, it is fundamentally limited by its disregard for spatial density. In regions of the manifold where units are sparsely distributed, k -NN may group perceptually distant sounds into the same bin simply because they



Figure 10.1: Sampling of an audio unit within a two-dimensional t-SNE projection of the RAVE latent space using the k -NN binning strategy. Discrete latent units are represented as points, with the reference unit (A) indicated by a blue arrow. The color gradient signifies the organization of units into discrete “bins” based on spatial proximity, where each bin corresponds to a specific value of the closeness parameter C .

are the closest relative to the rest of the corpus. The BD strategy addresses this by prioritizing the geometric topography over unit count. Instead of creating bins with an equal number of units, this algorithm divides the latent space into concentric categories of equal radial width based on a distance histogram. To do this, it utilizes *numpy*’s histogram function.¹

This approach, depicted in Figure 10.2, ensures that each bin represents a consistent degree of deviation from the original sonority. Consequently, bins in dense regions of the manifold will contain a high volume of potential candidates, while bins in sparse regions may remain empty. For the generated music, this results in significantly higher continuity at low values of C , as the algorithm restricts selection to a tight, absolute radius around the target. By making the sampling density-aware, the BD strategy maintains the morphological integrity of the original instrumental timbre more effectively than k -NN, particularly when navigating the specialized and often non-uniform clusters of a saxophone corpus.

10.2.3 Distance with Aligned Trajectory (BD-AT)

The BD-AT strategy further evolves the navigation logic by introducing the concept of angular displacement to account for the sequential nature of musical phrasing. In high-dimensional latent spaces, the relative direction between two consecutive units often encodes more information about

¹<https://numpy.org/doc/stable/reference/generated/numpy.histogram.html>, last accessed January 29, 2026.



Figure 10.2: Sampling of an audio unit within a two-dimensional t-SNE projection of the RAVE latent space using the BD binning strategy. Discrete latent units are represented as points, with the reference unit (A) indicated by a blue arrow. The color gradient signifies the organization of units into discrete bins based on spatial proximity, where each bin corresponds to a specific value of the closeness parameter C .

musical intent, such as the path of a specific gesture or a technical transition, than simple spatial distance alone. This algorithm operates by first calculating a reference vector between the current original audio unit and its immediate predecessor in the sequence of the score to establish a directional path for the reference sequence.

Once this reference trajectory is established, the algorithm estimates the angular distance between the original sequence vector and the vectors leading to potential units within a specific distance-based bin. By using a random distribution that is weighted according to this angular alignment with respect to the original sequence, the system gives preference to units that preserve a comparable directional flow, as illustrated in Figure 10.3. It ensures that even when the system is set to higher levels of divergence, the resulting sequence of audio units preserves a logical motion that remains understandable and coherent to the listener. The BD-AT strategy thus effectively mitigates the disjointed quality often associated with random sampling, providing a more contextually aware variation that respects the directional intent of the performer’s gestural language.

10.3 Evaluation Methodology and Materials

To assess the practical efficacy of the previously described algorithmic strategies, a systematic evaluation was conducted using a two-pronged approach that combines subjective perceptual testing with objective similarity metrics. The study employed three navigation strategies (k -NN, BD,

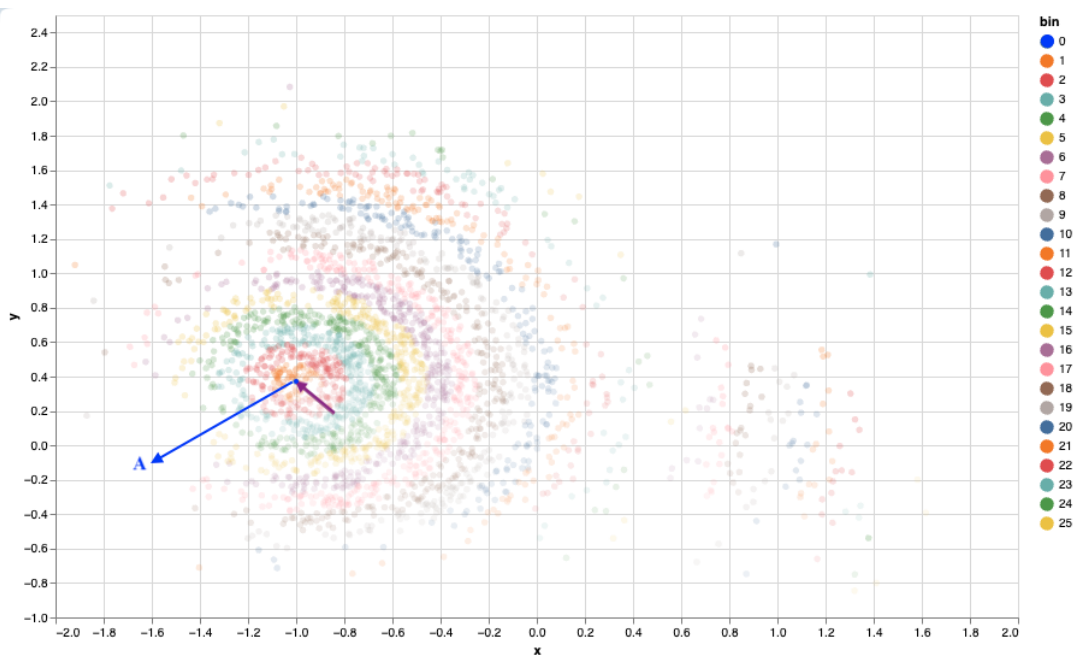


Figure 10.3: Sampling of an audio unit within a two-dimensional t-SNE projection of the RAVE latent space using the BD-AT binning strategy. Discrete latent units are represented as points, with the reference unit (A) indicated by a blue arrow. The color gradient signifies the organization of units into discrete bins based on spatial proximity, where each bin corresponds to a specific value of the closeness parameter C .

and BD-AT) to generate a diverse set of variations based on two distinct sections of the electronic tape from Jesús Villa-Rojo’s *Lamento*.

10.3.1 Context and Source Material: *Lamento*

The experimentation explores the navigation of the latent space derived from *Lamento* (1989) for tenor saxophone and tape (Villa-Rojo, 1989). Emerging from a collaboration between Villa-Rojo and the renowned saxophonist Daniel Kientzy (Kientzy, 1990), this work is particularly suitable for investigating timbre manipulation due to its free temporal structure and rich timbre nuances. While broader musicological surveys of Villa-Rojo’s catalog have documented his aesthetic evolution and integration of electroacoustics (Ordiz Castaño, 2008, 2011), *Lamento* remains largely unexplored within formal computational and generative analysis. Consequently, as demonstrated in our preliminary studies (Carvalho et al., 2025b), this gap presents an ideal opportunity to utilize the work as a novel case study for latent space navigation. The electronic tape itself is composed of pre-recorded saxophone and voice sounds. While the score allows the performer limited choices regarding speed and duration, they must synchronize with a clock to time specific musical phrases within fixed frames. These characteristics create an ideal landscape for exploring the relationship between human interpretive choices and machine-driven transformations.

To evaluate the efficacy of our sampling strategies, two contrasting sections of *Lamento*’s electronic tape were selected. The first section features a pre-recorded tenor saxophone playing

a sustained, long note characterized by a subtle vibrato, maintaining a consistent and stable timbre quality over its approximately 10-second duration. In contrast, the second section features a 10-second sequence where the saxophone executes a rapid alternation between two notes. This movement is further defined by a significant level of background reverberation, which causes the rapidly alternating frequencies to overlap and bleed into one another. In this context, this spectral overlapping creates a high degree of acoustic dissonance, manifesting as perceptual roughness and textural density rather than traditional harmonic tension.

These two distinct sections were deliberately chosen to test the robustness and behavior of the sampling algorithms under contrasting acoustic conditions as the closeness parameter (C) increases. Specifically, the first section serves to evaluate how well the system maintains morphological continuity and structural stability when processing a steady, sustained timbre. Conversely, the second section is intended to stress-test the trajectory alignment and density-based strategies against a highly complex, reverberant, and acoustically dissonant texture. By comparing the generated variations of these two extremes across discrete C values (from 0.04 to 0.96), the experiment seeks to determine which navigational logic most effectively preserves the original musical character, and at what threshold morphological drift causes the synthesis to lose cohesion.

10.3.2 Latent Space Generation: The RAVE Model

For this experiment, we leverage the model trained on the VocalSet dataset, which has demonstrated superior capabilities in clustering and recognizing tenor saxophone timbre and extended techniques compared to other generic models, as we discussed in Chapter 9. This model, trained by the Intelligent Instruments Lab (IIL), utilizes 16 latent dimensions and operates at a sampling rate of 48 kHz.

Despite its vocal training, the VocalSet model was selected for its demonstrated excellence in clustering and recognizing the timbre of tenor saxophones. It responds evenly across all pitch registers, except for dimension 16, which is more sensitive to the low register, and effectively handles dynamics and extended techniques. Its sensitivity to pitch and timbre variations provides a firm semantic basis for capturing the intricacies of the tenor saxophone.

10.3.3 Generation Parameters and Evaluation Framework

The effectiveness of these strategies was assessed through a multi-faceted analytical framework that combined subjective and objective metrics.

First, for each selected musical segment, three variations were generated across five discrete values of the closeness parameter C : 0.04, 0.2, 0.4, 0.6, and 0.96. To clarify the mechanics of this synthesis process, the generations rely on an AI-informed CBCS approach rather than synthesizing new waveforms from scratch via the RAVE decoder. In this framework, the three sampling strategies (k -NN, BD, and BD-AT) act as spatial querying algorithms that navigate the latent space to generate a sequence of indices. For this specific experiment, the audio source was segmented into grouped units of four latent frames, yielding discrete audio segments of approximately 170

milliseconds in length. Once the algorithms select a trajectory through the latent space, the corresponding raw audio segments are retrieved and sequentially concatenated via a Python script to construct the final audio variation. While the generations were constrained to the time frame of the source sections, the entire set of units related to the complete musical work was available for sampling. In all experimental runs, the latent space was organized into 25 bins, and all 16 dimensions were assigned equal weight for distance calculations.

The first evaluation component involved a perceptual assessment conducted by an expert saxophonist with over a decade of experience. This evaluation consisted of a systematic listening test across the generated variations for all five discrete C values (0.04, 0.2, 0.4, 0.6, and 0.96). To ensure the rigor of a quantitative test, this qualitative analysis graded the generations based on three specific dimensions of interest: 1) morphological and temporal Continuity, tracking the frequency of unwanted pauses, rhythmic interruptions, or disjointed audio jumps; 2) timbre identity, observing the preservation or degradation of the saxophone’s spectral characteristics (e.g., vibrato and reverberant tail); and 3) acoustic dissonance, evaluating the overall musical flow as the sequence deviated from the target.

This perceptual assessment was followed by an objective similarity measure, where the generated audio sequences were mathematically compared to the reference sequence using the `audio-similarity` Python library.² Features analyzed within this framework included the zero-crossing rate (to measure signal noisiness and pitch preservation), rhythm (to track temporal texture), chroma (to evaluate specific pitch content), energy envelope (to measure dynamic morphology and articulation), spectral contrast (to indicate perceived brightness and texture), and general perceptual similarity. By correlating these strict quantitative metrics with the expert listening dimensions, the framework provides a comprehensive understanding of the extent to which the generated audio retained the original’s acoustic characteristics under varying navigational logic.

This two-pronged approach, encompassing both subjective perception and objective measurement, enables a comprehensive understanding of how various latent space navigation strategies preserve or transform desired musical qualities. By comparing different strategies at various C values, we can determine which methods most effectively maintain original musical characteristics and where they diverge, providing deep insight into the impact of navigation logic on synthesis. It is worth noting that a *Google Colab* notebook³ was utilized as the testing environment to implement and evaluate these strategies. This computational environment enabled the thorough analysis of each audio unit and the systematic observation of the potential outcomes it generated.

10.4 Results and Discussion

The following analysis synthesizes the perceptual observations and objective analytical metrics gathered across the three developed navigation strategies: k -NN, BD, and BD-AT. By evaluating

²<https://github.com/markstent/audio-similarity>, last accessed January 29, 2026.

³https://github.com/NadiaCarvalho/IJPA-LatentSpace/blob/main/Experiment1_Mechanisms.ipynb, last accessed January 29, 2026.

these methods through both subjective listening and quantitative feature extraction, we can observe how the latent space manifold reacts to different sampling logics.

10.4.1 Perceptual Analysis of Sectional Variations

The first section analyzed features a sustained tenor saxophone long note with subtle vibrato, lasting approximately 10 seconds. When navigating this segment with the k -NN-based binning approach, the results at low C values ($C = 0.04$) closely resembled the original electronic version, maintaining musical flow with minimal dissonance. However, as C increased, the long note became subject to noticeable interruptions, pauses, and dynamic variations, resulting in a less cohesive listening experience despite remaining recognizable.

The BD strategy provided even greater continuity than k -NN for lower C values, often yielding results identical to the reference sequence due to the sparse population of units in the immediate vicinity of the original frame. However, this strategy exhibited significant sensitivity at higher C values, with several interruptions appearing as early as $C = 0.8$ because fewer latent units were available in the distant geometric bins.

BD-AT did not significantly improve low- C performance (i.e., the system’s ability to maintain high acoustic fidelity and morphological similarity to the original audio when the divergence parameter C is set near zero). In fact, these variations felt further from the source. We speculate that this phenomenon occurs because of the added constraint of angular alignment. By prioritizing units that follow the original sequence’s trajectory, the algorithm may overlook the absolutely closest unit in the latent space in favor of one that is spatially further away but more directionally accurate. While this sacrifice lowers raw similarity scores at near-zero divergence, it is the exact mechanism that prevents the incoherence observed in standard sampling at higher divergence levels ($C > 0.60$). Yet, at higher C values, the BD-AT strategy produced music that was markedly more discernible and coherent, despite interruptions. It establishes a logical flow that is almost opposite to the original, aligning better with the creative intent behind high-divergence parameters.

The second section, characterized by rapid note alternations and dense reverberation, showed similar trends. For k -NN-based binning, the furthest generations unexpectedly resembled sections with less movement. Distance-based binning at $C = 0.04$ produced sequences nearly identical to the original, maintaining fewer interruptions than the first section, even at higher values. Conversely, trajectory alignment in this dissonant context resulted in more significant interruptions as C increased, likely due to the complexity of the reverberant tail influencing the directional vectors in the high-dimensional space.

10.4.2 Comparative Analysis of Density and Temporal Continuity

The analytical results in Table 10.1 reveal a striking outlier: at the lowest divergence level ($C = 0.04$), the BD strategy achieves a rhythmic similarity of 0.7174, significantly outperforming both

k -NN (0.5333) and BD-AT (0.5492). This discrepancy highlights the fundamental difference between count-based and geometry-based sampling. Because the BD strategy utilizes an absolute distance histogram, a low C value restricts the sampling radius to a very narrow geometric shell. In the context of a specialized instrumental corpus like *Lamento*, latent units representing a continuous phrase are often clustered extremely close to one another. By enforcing a strict distance threshold rather than a fixed number of neighbors, the BD strategy ensures that only units within the immediate perceptual neighborhood are selected, effectively preventing the inclusion of units that may represent a different articulation or rhythmic onset.

Table 10.1: Mean similarity metrics for latent space navigation strategies. Higher values indicate greater preservation of original musical features (zcr, rhythm, chroma, energy, spectral contrast, and perceptual) relative to the source corpus.

Strategy	C	Similarity					
		zcr	rhythm	chroma	energy env.	spectral cont.	perceptual
k-NN	0.04	0.9989	0.5333	0.8988	0.5585	0.9266	0.4886
	0.20	0.9980	0.5449	0.8602	0.5643	0.9239	0.5125
	0.40	0.9976	0.5381	0.8433	0.5490	0.9187	0.4946
	0.60	0.9970	0.5381	0.8305	0.5685	0.9198	0.5157
	0.80	0.9959	0.5290	0.8128	0.5712	0.9131	0.5163
	0.96	0.9831	0.5345	0.7645	0.5270	0.9053	0.5205
BD	0.04	0.9997	0.7174	0.9778	0.7969	0.9585	0.7219
	0.20	0.9998	0.5317	0.9376	0.6186	0.9349	0.5375
	0.40	0.9988	0.5362	0.8952	0.5964	0.9286	0.5242
	0.60	0.9970	0.5379	0.8346	0.5640	0.9193	0.5007
	0.80	0.9916	0.5392	0.8006	0.5389	0.9115	0.5155
	0.96	0.9610	0.5306	0.7469	0.5061	0.9008	0.5046
BD-AT	0.04	0.9965	0.5492	0.8422	0.5842	0.9162	0.5345
	0.20	0.9972	0.5409	0.8429	0.5696	0.9200	0.5114
	0.40	0.9975	0.5356	0.8441	0.5761	0.9224	0.5402
	0.60	0.9971	0.5467	0.8396	0.5592	0.9198	0.5159
	0.80	0.9969	0.5238	0.7963	0.5483	0.9139	0.4949
	0.96	0.9964	0.5260	0.7485	0.5370	0.9121	0.4587

In contrast, the k -NN strategy is density-blind. Even at $C = 0.04$, it is still forced to select from the k closest neighbors. If the latent space is sparse in a particular region (common in electroacoustic music as it often employs many extended techniques), the k -th neighbor might be spatially distant, leading to an unintended jump in the audio signal. This distance explains the lower rhythmic and chroma scores for k -NN at low C values, as the algorithm pulls samples from across a wider perceptual gap to fill its bin quota. The BD-AT strategy also yields a lower score at $C = 0.04$ compared to the standard BD, due to the added constraint of angular alignment. By prioritizing units that follow the original sequence’s trajectory, the algorithm may overlook a closer unit in favor of one that is more directionally accurate. While this sacrifice slightly lowers raw

similarity scores at near-zero divergence, it is the exact mechanism that prevents the incoherence observed in standard sampling at higher divergence levels ($C > 0.60$).

10.4.2.1 General Trends in Feature Preservation

Beyond rhythmic continuity, the data provides valuable insights into the preservation and transformation of various musical elements. For all strategies, the zero-crossing rate (zcr) remains high across all C values, indicating a consistent preservation of pitch-related characteristics despite a slight decrease at the highest levels of divergence. This zcr result suggests that the RAVE latent space, particularly when using a model trained on vocal techniques, maintains a strong correlation with the fundamental periodicity and noisiness of the signal. Even when the system samples incorrect units from distant bins, the high-frequency content and periodic nature of the saxophone’s timbre (represented by the zcr) are robustly preserved, preventing the synthesis from devolving into uncorrelated noise.

The energy envelope similarity reveals how the system handles the dynamic morphology of the original phrases. While this metric remains relatively stable, the slight fluctuations observed (ranging from 0.50 to 0.79) indicate that higher C values introduce “jitter” into the amplitude profile. In the BD strategy at $C = 0.04$, the energy envelope similarity is at its peak (0.7969), confirming that tight geometric constraints preserve the original articulation and breath pressure patterns of the saxophone. As C increases, the introduction of units with different attack and decay characteristics causes the synthesized output to lose the smooth amplitude continuity of the original *Lamento* tape, though it rarely loses its overall gestural shape.

Similarly, the spectral contrast metrics, which evaluate the relative difference between peaks and valleys in the frequency spectrum, stay within a narrow, high range (0.90 to 0.95) for all strategies. This stability is significant because spectral contrast is a strong indicator of perceived brightness and texture. The high scores suggest that the RAVE model’s latent dimensions effectively disentangle the broad timbre texture of the saxophone from its specific pitch content. Consequently, even when the pitch content (chroma) varies significantly, the quality of the sound (e.g., its roughness, clarity, or metallic nature) remains consistent with the specialized language of the piece. These chroma values, representing specific pitch content, show the most significant decline across all strategies, reflecting the expected variability in pitch as the system explores the manifold.

In summary, lower values of the control parameter C effectively maintain the attributes of the original electroacoustic tape. While the BD strategy is most effective for high-fidelity replication at low C values, trajectory alignment (BD-AT) excels at establishing musical coherence during radical transformations. These findings suggest that a hybrid navigation approach could optimize results. However, the data also indicates that higher C values may fail to bring sufficient musical contrast, as core timbre and dynamic elements remain largely consistent across strategies.

10.5 Summary: Navigational Robustness and the Threshold of Coherence

The experimental analysis (Carvalho et al., 2025b) reveals that while all navigation strategies effectively preserve core musical elements such as timbre and dynamics, pitch elements demonstrate significantly greater variability as the control parameter C increases. Lower C values across all methods generally maintain the morphological integrity of the source material. However, as the sampling trajectory moves further from the target units into the periphery of the latent space, the resulting audio tends toward increased dissonance and rhythmic interruptions.

Notably, the incorporation of trajectory alignment (BD-AT) does not significantly alter musical outcomes at lower proximity levels. However, it proves superior for maintaining structural coherence during radical transformations, effectively mitigating the disjointed quality often associated with high-divergence sampling. These findings suggest that for complex instrumental timbres, such as the tenor saxophone, a hybrid approach combining distance-based categorization with trajectory awareness offers the most robust framework for meaningful navigation.

Despite the stability provided by these algorithmic mechanisms, a recurring challenge remains: the inherent difficulty in generating intentional musical contrast that respects a composer's stylistic logic or a performer's gestural intent. While the C parameter allows for sonic deviation, it does not explicitly account for the formal relationships of thematic development, such as the specific transformations required to create a contrary or oblique response to an input phrase. To address this, the following chapter introduces a formalized framework for navigating by musical motions, utilizing a specialized dataset designed to map these geometric relationships within the latent landscape.

Chapter 11

Mapping Gestural Intent: Taxonomies of Motion in Latent Space

The transition from purely proximity-based sampling to a system capable of intentional thematic development requires a formalized understanding of how musical ideas evolve in a digital latent space. Traditional corpus-based concatenative synthesis (CBCS) systems often struggle with this transition because they navigate multidimensional spaces blindly, relying heavily on step-by-step statistical matching or raw proximity targeting. Because these systems prioritize low-level acoustic similarity without semantic awareness, they lack compositional foresight, creating a profound disconnect between the machine’s retrieved units and the performer’s intended musical morphology. While earlier navigation strategies in our framework relied on mathematical distances, such as the Davies-Bouldin index, to measure cluster compactness, the creative demands of live performance require a more nuanced approach that mirrors human compositional thought. This chapter introduces a framework for motion-aware navigation, which departs from traditional point-to-point sampling by treating the latent space as a dynamic topography of musical gestures.

Central to this investigation is the Motiv dataset, a collection of expert tenor saxophone recordings explicitly annotated to illustrate the geometric behavior of thematic elaborations. By formalizing these elaborations through a taxonomy of musical motions (i.e., parallel, oblique, and contrary), this research seeks to bridge the gap between abstract neural activations and the structural logic of musical counterpoint.

The exploration presented here addresses a critical gap in current AI music systems: the ability to generate contrasting ideas while maintaining intentional structural coherence. By analyzing how professional performers interpret and execute these motions, we can map their unique trajectories within the RAVE latent space. This mapping enables the development of co-creative agents that do not merely mimic the performer but engage in a sophisticated dialogue through motion-driven retrieval and transformation.

Through the lens of the Motiv dataset, we move beyond viewing the latent space as a static database of sounds. Instead, we redefine it as a Motivic Landscape, where the relative positioning and directional flow of latent vectors provide the necessary data for contextually aware, expressive

music interaction.

The chapter is outlined as follows. Section 11.1 establishes a theoretical taxonomy of musical motions, categorized into parallel, similar, oblique, and contrary, to formalize the structural transformations and thematic development of musical phrases. Section 11.2 presents a preliminary experiment mapping these contrapuntal relationships within the RAVE latent space, evaluating how distinct performative gestures occupy unique geometric trajectories. Building upon these geometric signatures, Section 11.3 details the construction, curation, and structural organization of the Motiv dataset, a specialized corpus of expert tenor saxophone recordings explicitly designed to encode intentional performance gestures. Subsequently, Section 11.4 provides a preliminary mathematical analysis of this dataset, quantitatively assessing the structural correlation, directionality, and spatial variation of the recorded motion types. Finally, Section 11.5 synthesizes these findings, discussing their implications for transforming the latent space into a dynamic Motivic landscape and enabling contextually aware, motion-driven navigation in co-creative music systems.

11.1 A Taxonomy of Musical Motions

To introduce an interpretable structure into the latent space and guide the subsequent construction of the Motiv dataset, we adopt a taxonomy inspired by traditional two-voice counterpoint processes. While the term motion in multimodal AI often refers to human physical movement (Zhang et al., 2025), this research employs a distinct interpretation grounded in the work of Pinho (2016). Therefore, we separate the performer’s gesture (the physical and expressive execution, defined previously) from the musical motion. Here, motion refers to musical phrase elaborations that serve as structured transformations, shaping thematic development through specific geometric and performative relationships. This framework organizes musical transformations into four primary types, illustrated in Figure 11.1. These motions provide a foundation for identifying compositional patterns that enable a CBCS system to preserve a work’s stylistic context while dynamically aligning with a performer’s interpretation.

11.1.1 The Four Motion Types

The taxonomy categorizes the relationship between pre-existing musical material and the present musical discourse through the following definitions:

- **Parallel motion:** All musical voices or parameters move in the same direction by the same interval. In a performative context, this maintains the fundamental structure and contour of the original material, typically through direct pitch transposition with minimal deviation in expression.
- **Similar Motion:** The musical material moves in the same direction but by different intervals or magnitudes, retaining the broad properties of the source material while introducing minor variations in articulation, dynamics, or duration, allowing for stylistic modification.

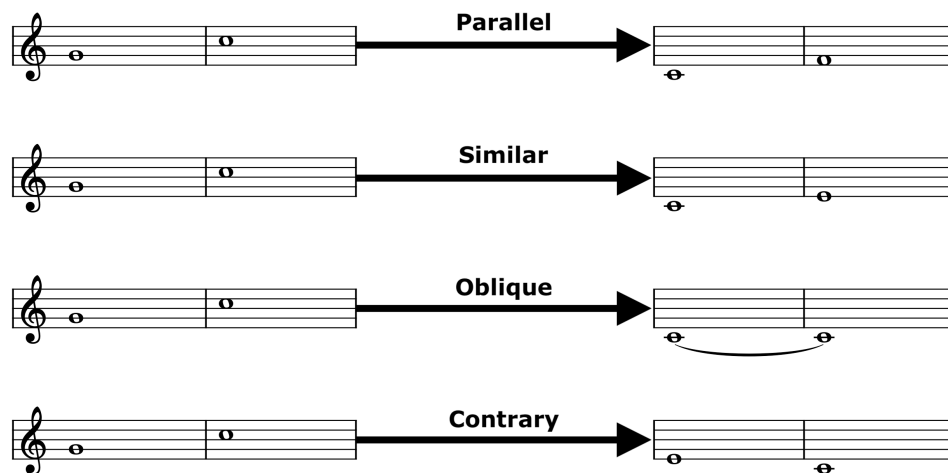


Figure 11.1: Illustration of the four motions within Pinho's taxonomy (Pinho, 2016). Left-to-right flow indicates the motion from pre-existing musical material to the present manifestation of musical discourse.

- **Oblique motion:** At least one core element remains stationary or preserved (e.g., pitch or rhythm) while others, such as dynamics, timbre, or context, are substantially modified. Thus, alters the perception of the elements without directly changing them, representing a selective transformation.
- **Contrary motion:** This represents the maximum deviation and expressive contrast from the original phrase. The musical parameters move in opposite directions, such as reversing interval directions, inverting dynamic envelopes (crescendo to decrescendo), or altering note durations to create stark opposition.

11.1.2 Consolidating the Taxonomy for the Motiv Dataset

For the practical implementation of the Motiv dataset and the preceding experiments, we opted to consolidate the taxonomy into three operational categories: parallel, oblique, and contrary. This decision stems from the observation that, in the context of live instrumental performance and latent space mapping, the distinction between parallel and similar motions often becomes blurred.

Our redefined parallel motion encompasses the conceptual space between Pinho (2016)'s original definitions of parallel and similar motions. In practice, a performer rarely achieves a mathematically perfect parallel transposition. Subtle fluctuations in timbre, microtiming, and expressive intensity naturally introduce the minor variations characteristic of similar motion. By merging these into a single parallel category, we acknowledge that any transformation maintaining the general contour and direction of the original phrase, even with slight intervallic or expressive deviations, occupies a consistent region of the latent space.

This consolidation simplifies the dataset’s architecture, focusing on the three most distinct geometric behaviors: maintaining identity (parallel), selective transformation (oblique), and radical opposition (contrary).

11.1.3 Implications for Latent Space Navigation

This taxonomy serves as more than a descriptive tool, providing the geometric logic required to navigate high-dimensional manifolds. By grounding the Motiv dataset in these expert interpretations of musical equality, variation, and contrast, we establish a foundation for models that can generate conventional patterns and novel oppositions.

Drawing on the nuanced expertise of seasoned musicians, this structured approach enables a co-creative system to move beyond simple stochastic sampling. Instead, it allows for interpretable navigation, where a performer’s live gesture can be mapped to one of these three motions, triggering a contextually aware response from the synthesis engine. This interpretable navigation pushes the boundaries of generative AI by ensuring that the latent space trajectories remain musically meaningful and compositionally relevant.

11.2 Experiment: Mapping Geometries of Motion

Before detailing the construction of the Motiv dataset used to train these behaviors, we examine the preliminary experiment conducted to evaluate how these motions manifest within a high-dimensional manifold.

This experiment seeks to map geometric equivalents of traditional contrapuntal relationships to determine if specific vector transformations in the RAVE latent space can reliably produce the four types of musical motion detailed in Section 11.1. By applying this taxonomy, we categorize parallel, contrary, and oblique motions to identify repeatable trajectories, effectively introducing interpretable control dimensions into the latent space. The primary objective is to evaluate how real-time saxophone gestures can manipulate these parameters, facilitating an intuitive and expressive dialogue between the performer and the synthesis system.

11.2.1 Methodology: Recording Gestural Variations

The experimental framework was conducted in two distinct stages, utilizing both saxophone-based and electronic-based phrases to explore the manifestation of musical motions. To construct this dataset, specific musical phrases were extracted from [Villa-Rojo \(1989\)](#)’s *Lamento* for tenor saxophone and tape. The recordings were performed by expert saxophonist Jorge Sousa, who was tasked with playing the original three excerpts selected from the written saxophone score and three replicated from the electronic tape, followed by a series of deliberate variations for each. These variations were explicitly performed to embody the parallel, oblique, and contrary motions defined in our taxonomy. Recordings and musical scores for each combination are documented in the project dataset ([Carvalho and Sousa, 2025a](#)).

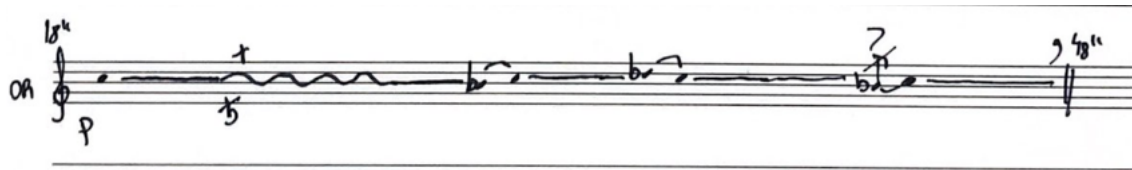


Figure 11.2: Musical score of phrase 1 from the saxophone part of Villa-Rojo (1989)'s *Lamento*, corresponding to 0:18–0:38.

11.2.1.1 Recordings of Motions in the Saxophone Score

In the first stage, three specific phrases from the saxophone part of *Lamento*, each lasting between 15 and 20 seconds, were selected for their distinct musical content and the varied emotional and gestural intents they convey, providing a comprehensive spectrum for latent space mapping.

The first phrase (spanning seconds 0:18 – 0:38) represents the saxophonist's initial entrance, following 18 seconds of solo tape, as illustrated in Figure 11.2. The melodic line is characterized by a static focus, hovering around a single note while maintaining a restricted, lower dynamic range. This phrase serves as a baseline for examining stable, low-divergence movement within the manifold.

In contrast, the second phrase (minutes 3:30 – 3:48), shown in Figure 11.3, introduces significant urgency and complexity. The melodic line oscillates around middle C before concluding with a rapid ornamental figure around a high A-flat at a forte dynamic. The sharp contrasts between piano and forte, coupled with shifts in register and articulation, provide a dense set of features for analyzing how the RAVE model clusters dynamic intensity and ornamental gestures.

The third phrase (minutes 7:30 – 7:45) focuses on temporal density and aggressive articulation. As seen in Figure 11.4, while the pitch material remains somewhat centered, similarly to the second phrase, the note transitions are significantly faster and include a distinct slap note. This phrase is crucial for evaluating the latent space's ability to differentiate between standard and extended techniques under conditions of high rhythmic velocity.

For each of these three phrases, the saxophonist recorded the original score as written, followed by a series of variations designed to embody the taxonomy of musical motions. Parallel variations involved simple transformation, such as transposing the phrases at different intervals while strictly adhering to the original rhythm and contour. Oblique variations maintained the pitch sequence but introduced shifts in dynamic envelopes and articulation. Finally, contrary motions were recorded by modifying or reversing the dynamics, extended techniques, melody direction, and accentuation

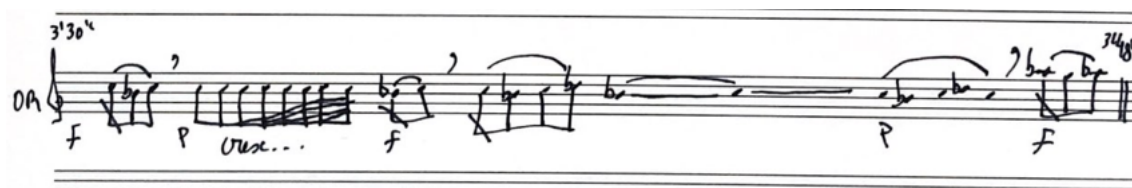


Figure 11.3: Musical score of phrase 2 from the saxophone part of Villa-Rojo (1989)'s *Lamento*, corresponding to 3:30–3:48.

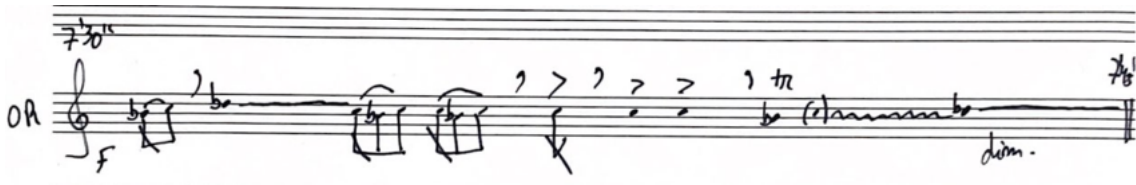


Figure 11.4: Musical score of phrase 2 from the saxophone part of [Villa-Rojo \(1989\)](#)'s *Lamento*, corresponding to 7:30–7:45.

patterns, such as transforming a crescendo into a decrescendo, effectively creating an inverted gestural path.

11.2.1.2 Recordings of Motions in the Tape Electronics

The second stage of the project focused on replicating and varying the tape electronics. Four short phrases were extracted from the tape, chosen for their unique musical characters and their potential to interact effectively with a live performer. The saxophonist was tasked with creating a mimic version of each, striving to replicate the timbre, dynamics, and phrasing of the electronic recording as closely as possible. Following this replication, the performer recorded two variations for each taxonomic motion (parallel, oblique, and contrary), specifically focusing on the attributes of the electronic sound rather than the written saxophone part.

11.2.2 Analysis: Latent Projections and Dimension Weighting

To analyze the recordings, we processed the audio data using the RAVE model, which was trained on the VocalSet dataset (as discussed in previous chapters). This process allowed us to extract latent space values for each audio unit. Consequently, the temporal audio segments were transformed into a sequence of points in a 16-dimensional space. We then projected these high-dimensional values into a two-dimensional representation using t-distributed Stochastic Neighbor Embedding (t-SNE) ([van der Maaten and Hinton, 2008](#)). This specific reduction technique was chosen over linear methods like Principal Component Analysis (PCA) because t-SNE excels at preserving local similarities, ensuring that audio units with similar timbre or spectral characteristics remain close to one another in the projection, making it possible to visualize and identify global trajectories and patterns within the otherwise abstract latent landscape.

By inspecting these visual paths, we established a clear connection between the performer's gestural intent and the resulting movement within the latent space. For instance, a crescendo or a shift in register manifests as a distinct directional trail in the t-SNE plot, allowing us to map musical motions to specific geometric vectors. This mapping provides the necessary groundwork for more effective manual or automated navigation during live performances, as it enables the system to recognize a performer's musical motion and respond with a variation that follows a complementary latent path. Furthermore, this analysis enabled the identification of latent anchors (i.e., regions of the space consistently associated with specific saxophone techniques), which serve as a vital reference for the weight attribution applied to the dimensions during distance calculations.

Building on the global trajectories observed in the t-SNE projections, we further refined our analysis by examining individual latent dimensions. As discussed in the Chapter 9, individual dimensions in the VocalSet-trained RAVE model exhibit high sensitivity to specific acoustic features of the saxophone. Dimensions zero and five are primarily linked to dynamics and general sound recognition, while dimension 10 is associated with pitch registers, and dimension 15 is connected to the recognition of extended techniques. By isolating these dimensions, we sought to determine whether the performer’s intended musical motions corresponded to shifts in the expected latent coordinates.

11.2.3 Results and Discussion

The results of this experiment, compiled in (Carvalho and Sousa, 2025a), align around consistent qualitative considerations. The motion trajectories reveal meaningful insights into how different gestural strategies inhabit the latent landscape. We present two primary examples, one from the saxophone score and one from the electronic tape, to illustrate these findings.

11.2.3.1 Analysis of Saxophone-Based Motions

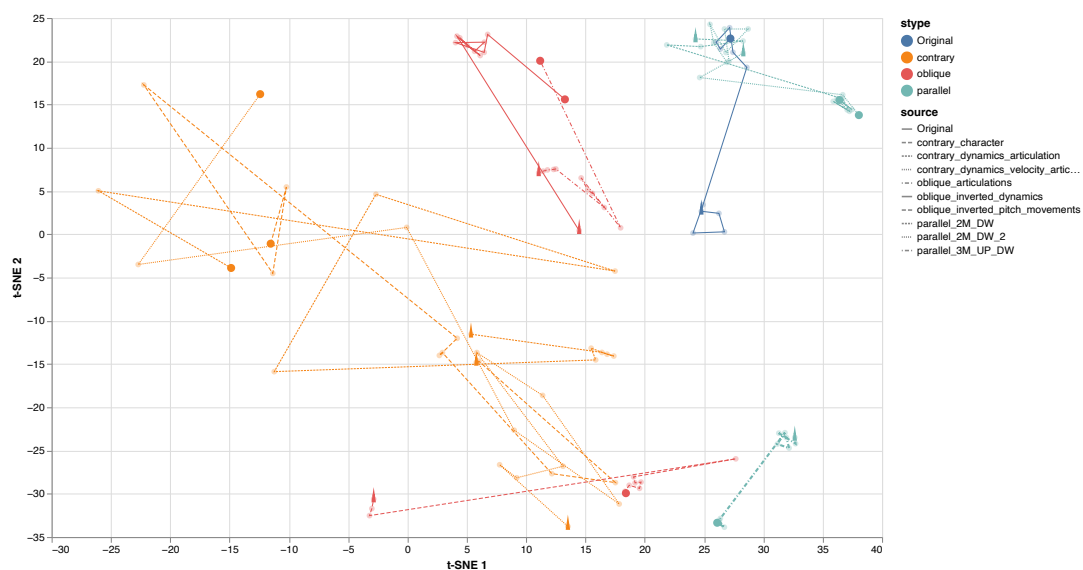


Figure 11.5: Two-dimensional t-SNE projection of RAVE latent trajectories (units 50–60) for the second saxophone phrase. The visualization illustrates the spatial dilation of contrary and oblique motions compared to the overlapping pathways of parallel transpositions. Motion types are color-coded with start points (circles) and endpoints (wedges) indicating temporal direction within the manifold.

Figure 11.5 displays an excerpt from the second saxophone phrase (units 50–60). In this global t-SNE projection, parallel trajectories tend to occupy the same vector length and configuration as the original phrase, though often in different locations or orientations within the manifold. For

instance, two parallel phrases undergoing a major second downward transposition nearly overlap, intersecting with the original phrase's path. Conversely, a major third transposition, while maintaining a similar shape, occupies a distinctly different spatial region.

Contrary and oblique motions exhibit shared behaviors, both resembling the original phrase's contour but with significantly more spatial dilation. Contrary motions show the most pronounced expansion, while oblique motions are less dilated. The distinction between these two is often nuanced; contrary motion may involve inverted dynamics or techniques, while oblique motion preserves specific elements while radically altering others. Because the performative stimuli for these motions are often related, they manifest as comparable spatial patterns that occupy areas further removed from the original phrase than parallel variations.

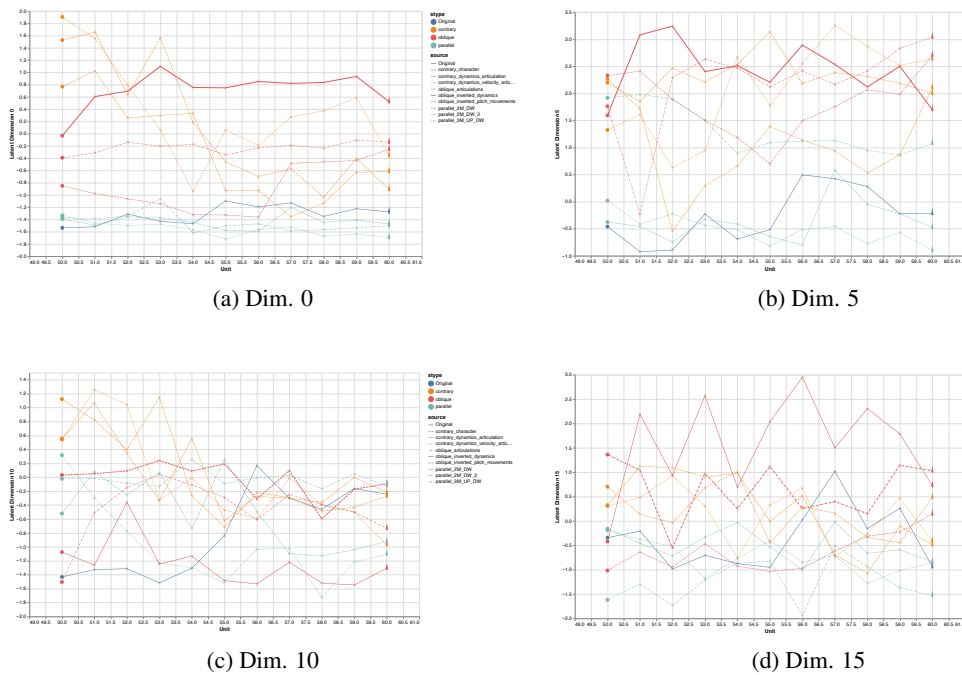


Figure 11.6: Latent space trajectories (dims 0, 5, 10, 15) for saxophone phrase 2. The comparison highlights structural consistency in most dimensions, alongside pitch-sensitive variance in dimension 10. Directional markers (circles to wedges) and color-coding track temporal flow and motion types across units 50–60.

When examining specific latent dimensions linked to the different qualities of the saxophone's sound (Carvalho et al., 2024), we can draw similar conclusions. In Figure 11.6, we can see dimensions 0, 5, 10, and 15 for the second saxophone phrase and its variations. In every dimension except the tenth, the parallel motion closely mirrors the original score, both in the paths and the space it occupies.

Intriguingly, this is not the case in the latent dimension 10, associated with pitch, as the parallel motion variations follow the original score with exact pitch transposition but deviate spatially. It appears that this latent dimension is exceptionally sensitive to pitch, capturing very subtle changes.

Consequently, the trajectory of these motions is not as similar to the original as in the dynamic-focused dimensions. This difference in latent dimensions' specific trajectories supports our decision to assign different weights to each latent dimension for calculating distances. This way, we can fine-tune motions according to specific structural properties of the saxophone performance and the intention behind each motion.

In Figure 11.6, oblique and contrary motions present similar behaviors across all individual dimensions. The contrary motion is generally more dilated across the whole space. It tends to follow the original trajectory's general configuration less strictly. In contrast, oblique motion follows particular segments of the configuration more closely but remains further from the original in the manifold compared to parallel motion.

11.2.3.2 Analysis of Electronic-Based Motions

In the second stage of the experiment, variations were derived from the electronic tape phrases (Figures 11.7 and 11.8). A primary observation is that the original electronic audio units are consistently isolated from all other recordings, including the saxophone mimic phrases. This spatial distance is likely influenced by recording conditions, post-processing, and the inherent density of the mastered tape, which includes added reverberation and layered saxophone sounds.

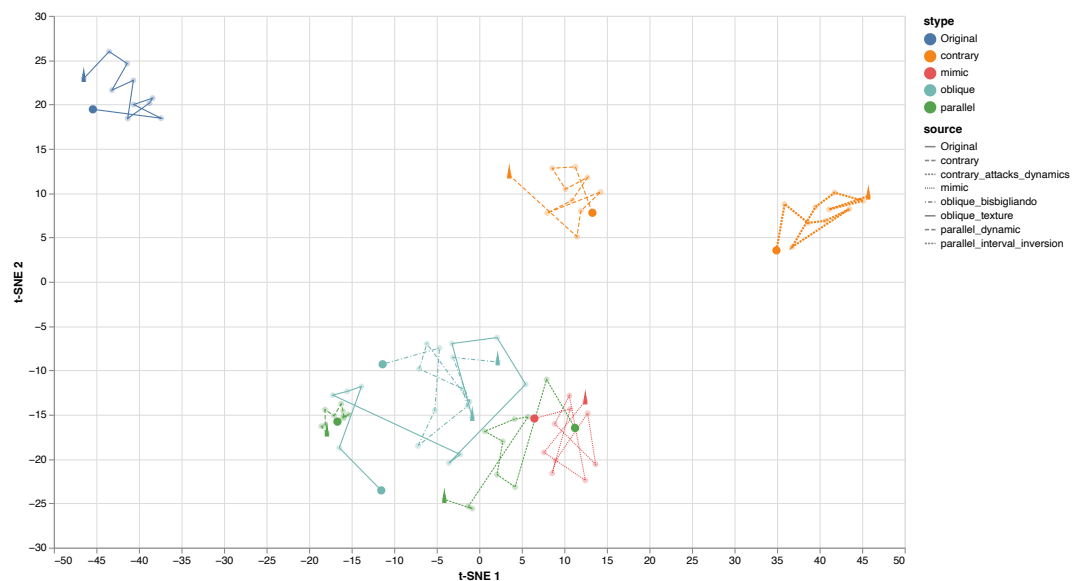


Figure 11.7: Global t-SNE projection of latent trajectories (units 50–60) for the first electronic phrase. The visualization highlights the significant spatial isolation of the original electronic source from all saxophone-based variations and mimics. Parallel, oblique, and contrary motions are color-coded to illustrate the varying degrees of spatial dilation and directional alignment relative to the saxophone mimic.

Consistent with the first stage, Figure 11.7 shows that contrary motions exhibit the most significant dilation. Parallel motions again mirror the mimic phrase in spatial occupation, following

similar vector shapes. Interestingly, in this context, oblique motions cover more total space than contrary ones, though they remain closer to the mimic and parallel phrases.

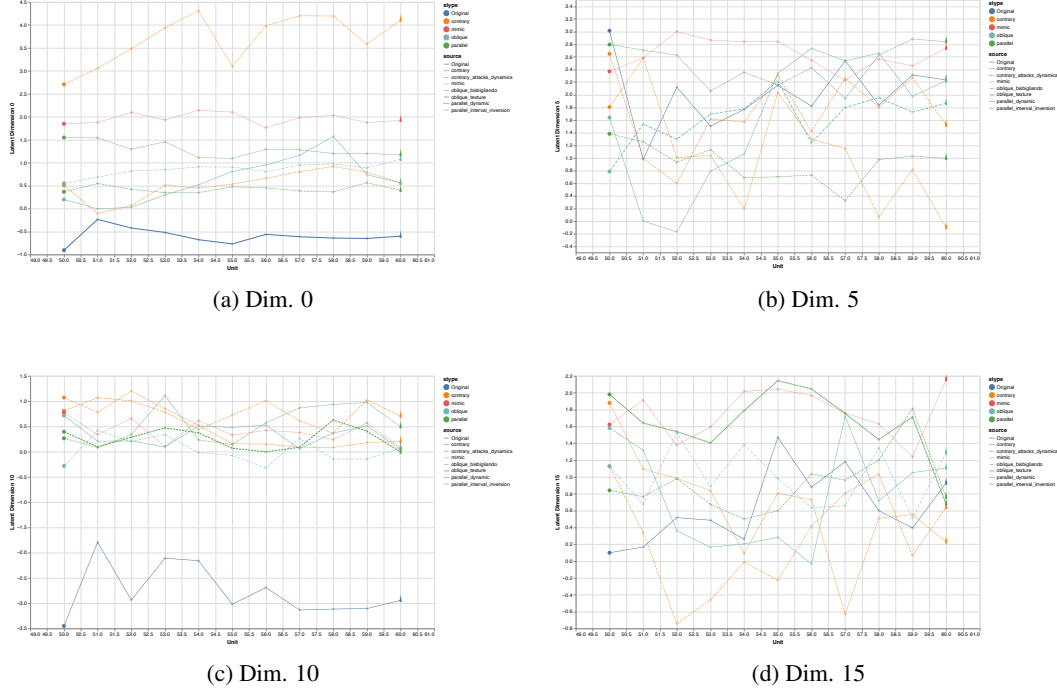


Figure 11.8: Individual latent trajectories for dimensions 0, 5, 10, and 15 (units 50–60) of the first electronic phrase. The plots reveal a distinct spatial offset in dimension 10, indicating the latent sensitivity to recording density and tape mastering versus live saxophone replication. In dimensions 5 and 15, parallel motions demonstrate high directional correlation with the electronic source. In contrast, contrary motions exhibit significant divergence in the technical register (dim. 15). Temporal sequence is indicated by filled circles (start) and upward wedges (end).

Analysis of individual dimensions for the electronic phrases reveals that dimension 10 shows the most significant gap between the electronic source and the saxophone replications. We hypothesize that this dimension is sensitive to signal density, such as the layering of multiple sounds found in the tape, which a live saxophone cannot replicate. In dimensions 5 and 15, however, parallel motions closely follow the electronics. In dimension 15 specifically, contrary motions move in the opposite direction to the electronics, a characteristic that could be leveraged within the system to create intentional musical tension or divergence.

11.2.4 Synthesis: From Geometric Mapping to Gestural Instruction

The results of this experiment (Carvalho et al., 2025b) provide empirical evidence that the taxonomy of musical motions (parallel, oblique, and contrary) manifests as distinct geometric signatures within the latent space. The visualization of these trajectories confirms that while the RAVE model can preserve the general shape of a musical gesture, it does so with varying degrees of sensitivity across its dimensions. Specifically, the observed spatial dilation in contrary motions and the

high sensitivity of dimension 10 to pitch transposition highlight the necessity of a more nuanced navigational approach than simple Euclidean distance.

The acoustic gap identified between the electronic tape and the live saxophone mimicry further suggests that the latent space is highly sensitive to production-level features, such as recording density and reverberation. This finding is crucial, as it implies that for a co-creative system to interact meaningfully with a live performer, the training material must be specifically curated to reflect the instrument’s gestural possibilities, rather than relying on mastered studio recordings.

Furthermore, the directional correlation observed in dimensions 5 and 15 provides a clear path for implementing motion-based weight attribution. By prioritizing these dimensions, our system can more effectively follow the performer’s trajectory or intentionally diverge from it, transitioning from a reactive sampler to a contextually aware musical partner. These insights directly inform the construction of the Motiv dataset. Rather than focusing on a specific musical work or complex electronic textures, this specialized training corpus is designed to capture the fundamental motions of saxophone performance. By isolating these gestural pairs in the recording process, we provide the model with a dense, semantically organized vocabulary of instrumental motion, bridging the gap between abstract latent navigation and intentional musical dialogue.

11.3 The Motiv Dataset: Encoding Performance Gesture

The preliminary experiments confirmed that the taxonomy of musical motions (parallel, oblique, and contrary) manifests as distinct geometric signatures within the RAVE latent space. While the model preserves the general morphological shape of a gesture, it does so with non-uniform sensitivity across dimensions, such as the high sensitivity of dimension 10 to pitch transposition. Furthermore, the acoustic gap between electronic tape and live saxophone mimicry reveals that the latent space is highly susceptible to production features such as recording density and reverberation.

These findings suggest that for a co-creative system to achieve meaningful interaction, training materials must be specifically curated to reflect the instrument’s gestural possibilities, rather than relying on mastered studio recordings. To address this, we developed Motiv, a specialized dataset of expert saxophonist recordings designed to formalize the relationship between thematic elaborations and their latent representations. By isolating gestural pairs (parallel, oblique, and contrary) within a controlled recording process, we provide the model with a dense, semantically organized vocabulary of instrumental motion. This approach bridges the gap between abstract latent navigation and intentional musical dialogue, transitioning the system from a reactive sampler to a contextually aware musical partner.

As existing datasets primarily focus on motif repetition in the symbolic domain, they often lack the timbre nuance required for high-fidelity audio interaction. Motiv fills this gap by leveraging expert interpretations to provide a foundation for diverse, contextually aware compositions. The following sections detail the acquisition and exploratory analysis of this dataset, systematizing the geometric relationships that enable stylistic preservation alongside interpretive freedom.

11.3.1 Recording Protocol and Dataset Construction

Following the taxonomic consolidation, we developed a rigorous recording protocol for the Motiv dataset. The primary objective was to move beyond preliminary experiments by creating a dense, structured corpus of expert saxophone interpretations that explicitly map the relationships between an anchor phrase and its taxonomic variations.

The construction of the Motiv dataset followed a structured five-step process: 1) selecting expert saxophonists and determining the recording setup; 2) defining the source materials; 3) establishing a controlled set of parameters to represent parallel, similar, contrary, and oblique modes; 4) modeling the recorded musical phrases as vectors in a latent space; and 5) structuring the annotated dataset to ensure consistency and usability for further analysis. This procedure was designed to systematize musical phrase elaborations and evaluate their geometric manifestations within a high-dimensional manifold.

11.3.2 Setup and Participants

To ensure the creation of a high-quality corpus for latent space analysis, we recruited four professional male saxophonists, aged 23 to 34, all of whom hold master’s degrees in music. These performers were selected based on their proficiency in contemporary repertoire, specifically works like Villa-Rojo’s *Lamento*, and their extensive experience with mixed music for saxophone and electronics. Their professional background emphasizes nuanced interpretation and free improvisation, fostering a collaborative approach to developing new repertoire. Furthermore, as the tenor saxophone possesses a highly distinctive musical identity, consistent proficiency with this specific instrument was a critical selection factor.

The recordings were conducted in January 2025 within an auditorium setting. The technical setup utilized a stereo pair of Sontronics STC-1S MT BK microphones¹ and the REAPER digital audio workstation software², capturing audio at a 48 kHz sampling rate (later downsampled to 44.1 kHz at 24-bit resolution). To maintain strict temporal alignment between the thematic elaborations and the original score, synchronization was achieved via a chronometer visible to the saxophonists during their performance.

11.3.3 Dataset Materials and Guidelines

To ensure high-quality material for training and analysis, we recorded the four expert saxophonists performing three curated phrases from *Lamento*, as described in Section 11.2.1.1.

For each phrase, performers were asked to record the original version as written in the score, followed by two to four interpretations of parallel, oblique, and contrary motions, as defined in Section 11.1. While we provided conceptual guidelines for each motion based on Pinho’s taxonomy, the specific musical execution was left open-ended to capture the interpretative range and

¹<https://sontronics.com/products/sontronics-stc-1s-in-matched-stereo-pair-silver>, last accessed January 29, 2026.

²<https://www.reaper.fm>, last accessed January 29, 2026.

individual nuances of each performer. Preparation instructions containing the conceptual guidelines³ were shared with the saxophonists three weeks before the recording session.

11.3.3.1 Parallel Motion

These variations maintain the fundamental structure and contour of the original phrase. As shown in Figure 11.9, executions often relied on direct transpositions or subtle modifications in dynamics and articulation while preserving the overall character.

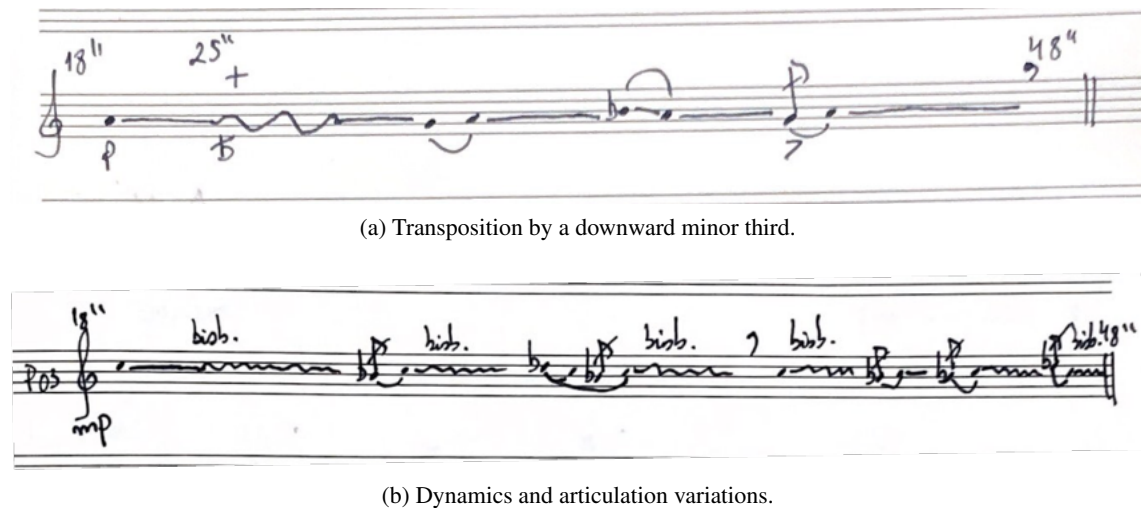


Figure 11.9: Examples of parallel motions applied to the saxophone score of phrase 1.

11.3.3.2 Oblique Motion

This transformation preserves at least one fundamental element while substantially modifying others, such as dynamics, articulation, or interval relationships. Examples in Figure 11.10 illustrate selective shifts, such as changing a dynamic from piano to forte or altering appoggiaturas.

11.3.3.3 Contrary Motion

Contrary motions focus on maximum deviation, reversing key elements like interval direction, note duration, and dynamic envelopes. Figure 11.11 demonstrates stark contrasts achieved through register shifts and radical timbre changes, such as the addition of trills or subtones.

11.3.4 Motion Distribution and Challenges

The expert interpretations employed a range of musical strategies to achieve these effects. Table 11.1 details the relative frequency of these motion categories within the dataset.

³https://github.com/NadiaCarvalho/Motiv-Dataset/blob/main/experiment_instructions.docx, last accessed January 29, 2026.

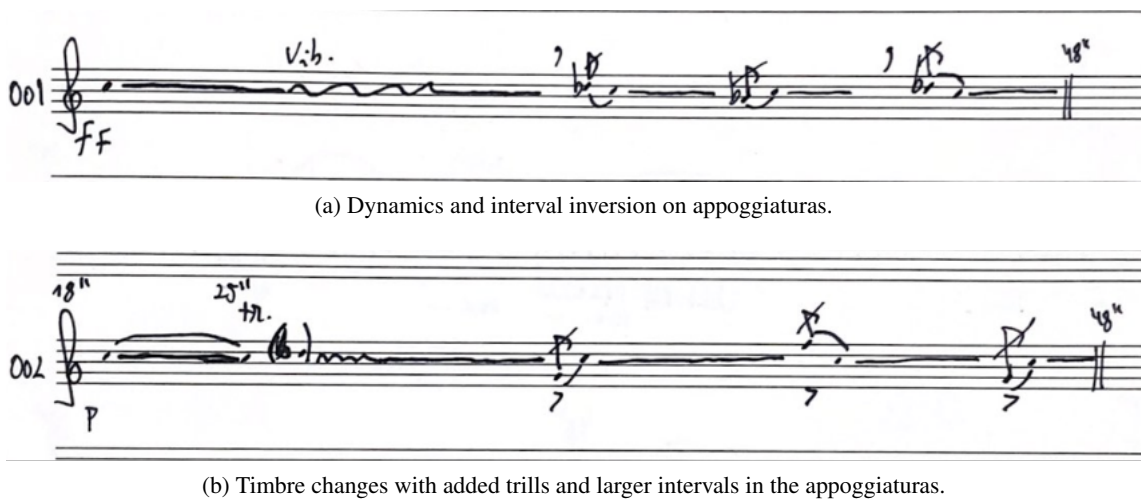


Figure 11.10: Examples of oblique motions applied to the saxophone score of phrase 1.

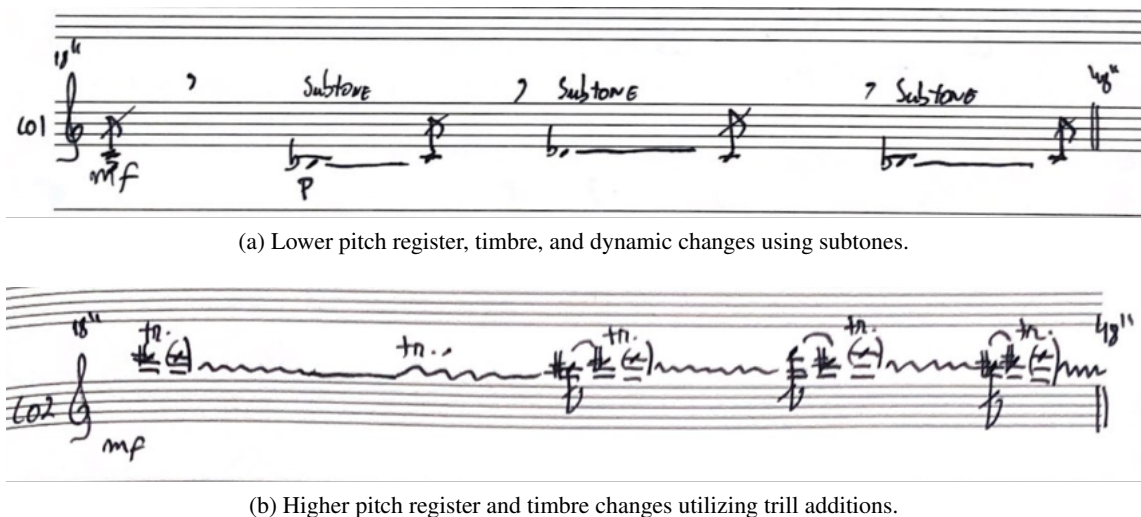


Figure 11.11: Examples of contrary motions applied to the saxophone score of phrase 1.

Analysis of these strategies reveals that pitch changes are the dominant tool for parallel motions (55%), while rhythmic changes (specifically phrase reversals) are the primary driver for contrary motions (50%). Performers noted that creating meaningful oblique motions was particularly challenging, often finding it difficult to execute transformations that were distinct yet structurally tied to the anchor. Consequently, the dataset contains fewer oblique samples compared to the more intuitive parallel and contrary types.

11.3.5 Data Modeling and HDF5 Organization

The final stage of the Motiv dataset construction involved modeling the recorded phrases within a latent space and structuring the resulting data for computational accessibility. To generate these representations, we utilized the RAVE model, which was trained on the VocalSet dataset. This specific model, as discussed in the previous two chapters, was selected due to its superior ability

Table 11.1: Percentage of musical samples exhibiting each transformation category within the Motiv dataset. The data illustrate the dominance of direct transposition in parallel motions (55%) and the prevalence of phrase reversals and dynamic inversions in contrary motions (50% and 30% respectively).

Transformation Category	Parallel (%)	Oblique (%)	Contrary (%)
Pitch Changes (PC)	55% – Direct transpositions (M3, P4, P8)	30% – Multiple interval shifts	15% – Interval inversions and phrase reversals
Rhythmic Changes (RC)	10% – Tempo	40% – Inverted duration and free changes	50% – Phrase reversal
Articulation & Dynamics (AD)	20% – Dynamics and transpositions	50% – Articulations and dynamics	30% – Dynamic inversions
Timbre & Effects (TE)	15% – Bisbigliando/trills and pitch	20% – Timbre, vibrato, oscillations	30% – Free timbre effects

to handle pitch registers, extended techniques, and dynamics simultaneously when applied to the tenor saxophone, outperforming models trained on instrumental improvisation.

The latent extraction process follows a structural logic where each audio recording is resampled to mono and projected into the RAVE encoder. The model compresses the input waveform into a sequence of 16-dimensional latent vectors that capture a complex mix of features related to timbre, pitch, articulation, and dynamics. For a one-second phrase at a 48 kHz sampling rate, the model produces 23 compressed audio frames, resulting in a 23×16 dimensional matrix of floating-point numbers, similar to the one shown in Figure 11.12. Although disentanglement is not explicitly enforced, specific dimensions correlate with interpretable attributes. For instance, dimension 16 is sensitive to the low register, while dimension 13 responds exclusively to the note C across all registers.

To ensure efficient management of audio, latent vectors, and metadata, the dataset is stored in HDF5 format. This format was adopted because of its efficiency in handling large-scale scientific data and annotations. As illustrated in Figure 11.13, the HDF5 file is organized into four main branches, facilitating seamless integration and analysis. The `audio_samples/` branch contains the waveform arrays for each recording. The `latent_vectors/` branch stores the $N \times 16$ arrays of RAVE latent values, where N remains consistent across samples of the phrase due to temporal synchronization. The `annotations/` branch includes textual descriptions of the performers’ intentions and specific modifications, such as transpositions or interval shifts. Finally, the `musical_scores/` branch provides references to image files of the scores for each variation.

Within this structure, the 116 recorded samples in the Motiv dataset are categorized using a systematic labeling convention based on the phrases, saxophone player, and motion type. Each

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	d13	d14	d15	d16
0	1.7404	-0.0409	0.2040	0.2115	-0.2893	1.2161	0.7856	-0.6710	-0.0371	0.1070	0.8015	-0.2241	-0.0456	3.0413	1.0178	-0.5402
1	1.7685	-0.7006	-0.7171	0.8807	0.2840	1.6246	-0.3624	-1.0461	-0.8566	0.4059	0.7080	0.4766	-0.9458	3.9926	0.5522	-0.3648
2	1.4692	0.1424	1.2849	-0.6518	-0.4973	0.5518	-0.5165	-0.6726	-0.1135	0.0723	0.3160	0.6763	-0.5987	-3.5251	-0.1263	2.7460
3	2.1644	-0.8501	1.1928	-0.1599	-0.6350	-0.6482	0.7227	0.3642	-0.5691	0.3514	0.1573	0.9097	1.2873	0.2471	-0.6473	5.2597
4	2.0572	0.7182	-1.2159	0.6027	0.2436	0.9789	-0.8163	-0.9520	0.6204	-0.7963	1.1875	-0.4989	0.5722	-0.7934	-0.3598	-2.1132
5	2.1069	-0.5292	0.9466	0.4590	1.0143	1.9386	-0.2884	-0.6942	0.8996	-0.0545	1.4771	-0.7851	0.3382	2.9436	-1.0422	0.1622
6	2.1816	-1.3801	-0.0688	0.7743	0.1931	1.6406	0.8305	-0.7545	-0.5543	0.6294	0.5810	-0.5270	0.9998	-1.9575	0.0042	1.1422
7	1.4983	1.4792	-0.5904	-0.5032	-0.4071	1.3221	-0.8229	-0.0289	-0.3903	0.8868	0.7329	-0.5431	-1.9868	-1.6558	-1.0653	-2.6024
8	1.7103	0.6022	-0.7688	-0.3157	-0.3626	1.7840	-0.4107	0.4490	-0.5737	-0.7750	1.0256	-1.4110	-0.8808	2.9085	0.5606	0.0769
9	1.9291	-0.4361	1.2904	-0.2824	0.4027	1.1818	1.3309	-0.8165	-1.4954	0.1622	-0.5622	0.9984	-0.0528	-1.6111	0.2204	3.3087
10	1.8229	-0.1831	-0.3554	-0.3679	0.1029	1.6526	0.0597	-0.4491	-1.1229	0.2517	0.8001	-0.5515	-0.4976	0.4124	1.5016	0.1036
11	2.0733	0.7679	0.1588	-0.2908	0.8462	2.5628	-1.0782	0.8060	-0.8690	-0.8903	1.3604	-0.8205	0.2811	-0.6055	0.5278	-0.2058
12	1.6918	0.7305	0.1513	-0.6371	0.5672	3.0778	0.1549	0.7172	-1.6342	0.7222	2.2250	0.2294	-0.4875	2.5091	0.0762	-1.1665
13	2.1418	-0.3506	-0.5426	-0.3441	-0.9896	1.8761	0.5081	-0.5686	0.2019	-0.0195	0.8826	-0.2729	-0.4938	-3.2330	-0.0074	0.2923
14	1.3119	2.2412	-0.8376	-0.5841	0.7980	2.7282	-0.4340	-0.7286	-0.0894	0.1904	0.7563	0.2541	-1.3028	-0.7189	-1.4860	0.3059
15	1.5377	-0.6775	0.4844	-0.3870	0.8944	1.6017	0.0100	0.2589	-1.5175	0.5111	-0.9860	0.4386	0.0914	2.0150	-1.1736	4.5663
16	2.0931	-1.6309	0.9583	-0.2873	-0.3744	1.7376	0.4243	-0.2418	-1.7162	0.9312	0.5144	2.3673	0.5891	1.2572	-1.0262	0.3399
17	1.3790	0.7940	0.0231	-0.6425	1.1135	2.0770	0.6575	0.5119	-0.5439	-1.1486	0.6929	0.6000	0.8807	-4.2466	-0.0339	-0.8043
18	1.6116	2.6636	1.2462	-0.4476	0.2543	2.2537	0.0956	-1.1584	-0.3151	-0.8840	0.2261	0.5021	0.2678	-1.1833	0.4124	1.3282
19	1.7725	1.1659	-1.0940	0.6500	1.2824	1.3421	0.5732	0.5991	-0.8169	-0.4415	1.8309	-0.0309	-0.8285	5.5319	-0.0914	0.4453
20	1.4571	-1.4364	-0.4016	0.3894	0.6979	1.8834	0.4541	-0.1322	-0.3719	1.3356	1.0424	0.9348	-1.6880	1.8954	1.2321	0.5714
21	2.0419	-2.0586	0.4105	0.0444	0.6864	2.5067	-0.3574	0.4074	0.0622	0.7210	1.1855	0.1487	0.5823	-0.8117	0.3512	0.4397
22	1.7544	1.7462	-0.0849	-0.9670	0.4518	2.0736	0.5778	1.6256	-0.5245	1.4716	0.5162	0.0960	-0.9217	-3.3772	-0.6532	-1.1564

Figure 11.12: Latent vectors corresponding to a phrase lasting one second result in 23 compressed audio frames, leading to a latent matrix that is a 23 by 16-dimensional vector containing floating-point numbers. Dimensions in the figure are labeled d1 to d16, with timesteps ranging from 0 to 22.

phrase is designated as PH followed by a number from 01 to 03 to represent its source phrase. After the phrase number, the performer is encoded with an “S” followed by a performer number (01 to 04) to identify the specific saxophonist. Finally, the type of motion is denoted by OR for original, “P” for parallel, “O” for oblique, or “C” for contrary, with a trailing number indicating the recording order. For example, the label PH01_S01_P01 represents Saxophonist 01’s first parallel motion variation of the first phrase.

To enhance the dataset’s utility, players provided detailed annotations regarding their expressive intentions and specific modifications, such as transposing the original phrase by a major third (sample PH01_S01_P01). Furthermore, individual musical scores were included for every recorded sample to support comparative performance analysis. This comprehensive organization ensures that the Motiv dataset (Carvalho and Sousa, 2025b) is robustly prepared for diverse Music Information Retrieval (MIR) tasks, including gesture-based recognition, motion-driven synthesis, and the exploration of geometric relationships in latent space.

```

Motiv.h5
├── audio_samples/
│   ├── PH01_S01_OR (waveform array)
│   ├── PH01_S01_P01 (waveform array)
│   ├── PH01_S01_P02 (waveform array)
│   └── ⋮
├── latent_vectors/
│   ├── PH01_S01_OR (Nx16 array, where N = frames)
│   ├── PH01_S01_P01 (Nx16 array)
│   ├── PH01_S01_P02 (Nx16 array)
│   └── ⋮
├── annotations/
│   ├── PH01_S01_OR ("Original phrase 1")
│   ├── PH01_S01_P01 ("Transposition by
│   │   upward major 3rd")
│   ├── PH01_S01_P02 ("Transposition by
│   │   downward minor 3rd")
│   └── ⋮
└── musical_scores/
    ├── PH01_S01_OR.png
    ├── PH01_S01_P01.png
    ├── PH01_S01_P02.png
    └── ⋮

```

Figure 11.13: Structural organization of the Motiv dataset.

11.4 Preliminary Dataset Analysis

The preliminary analysis of the Motiv dataset sought to determine if the consolidated taxonomy of musical motions exhibits consistent geometric relationships within the latent space relative to the original phrases. To investigate this, we calculated four specific metrics for each pair of recordings to analyze three distinct geometrical characteristics: structural correlation, directionality, and variation. The following sections will describe the experimental methods, report the results, and provide a discussion of their implications. To facilitate reproducibility, the code used to compute the preliminary analysis has been made available online as a *Google Colab* notebook.⁴

⁴<https://github.com/NadiaCarvalho/IJPA-LatentSpace/blob/main/PhraseRecordingsAnalysis.ipynb>, last accessed January 29, 2026.

11.4.1 Experimental Methodology

The experimental methodology focused on directly comparing the latent trajectories of these recorded variations with those of the original phrases performed by the same saxophonist. This approach allowed us to normalize the data by overlooking significant acoustic changes related to individual saxophone setups, such as mouthpieces and reeds, and focusing strictly on the performative gesture. We hypothesized that systematic geometric shifts would exist between motion types, which were assessed through three distinct analytical lenses: structural correlation, directionality, and spatial variation.

To ensure that the analysis captured meaningful audio characteristics rather than single-frame noise, we processed the latent vectors in windows of four frames, calculating the average values for each dimension within those windows.

11.4.1.1 Structural Correlation

Structural correlation is crucial for understanding how temporal sequences, such as audio trajectories, evolve and interact with one another. We measured this using the Pearson Correlation Coefficient (PCC), which identifies linear relationships between corresponding dimensions of two sets. High PCC values indicate shared linear patterns and structural similarity, whereas values near zero suggest the absence of a linear relationship. The coefficient r is defined as:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}, \quad (11.1)$$

In this equation, the numerator represents the covariance between the sets, while the denominator normalizes the result by their standard deviations.

11.4.1.2 Directionality

Directionality refers to how sequences develop within the learned representation, allowing us to comprehend the temporal flow and consistency. Directionality was evaluated using two complementary measures:

- **Angular distance to the centroid:** This scale-independent metric evaluates how much a movement deviates from the average trajectory. By focusing on the centroid, which represents the average position of a trajectory, we can compare motion patterns while reducing the influence of variations in magnitude. This criterion is especially useful in music, where performers may alter their intensity without changing their underlying expressive intent.
- **Cosine similarity:** This metric measures how closely different variations or performers align within the latent space by calculating the cosine of the angle between two vectors. A high similarity indicates a shared directional trend, while low similarity reveals structural deviations or individual differences among performers.

11.4.1.3 Variation

Variation reflects the absolute differences caused by expressive performance or structural changes. We quantified this variation using Euclidean Distance, which measures the absolute deviation between corresponding time points in the latent space. Larger distances indicate greater variation from the original performance. The distance d between two points p and q is calculated as:

$$d(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (11.2)$$

11.4.2 Results and Discussion

The following results clarify the geometric relationships between the motion types and the original phrases, validating the hypothesis that these transformations occupy distinct regions and trajectories within the latent space.

11.4.2.1 Structural Correlation

The structural correlation of latent time points provides insight into how temporal sequences evolve and relate to one another. Table 11.2 presents the PCC values comparing various motion types to original phrases across all performers. This metric identifies the degree of structural alignment between recorded elaborations and the source material.

Table 11.2: Pearson Correlation Coefficients (PCC) between recorded motion types and original phrases across four saxophonists. Unless otherwise noted with * ($p < 0.05$) or † (not significant, $p > 0.05$), all values indicate $p < 0.0001$. Higher values reflect stronger structural similarity in the latent space.

	Saxophonist 01			Saxophonist 02			Saxophonist 03			Saxophonist 04		
	Ph 1	Ph 2	Ph 3	Ph 1	Ph 2	Ph 3	Ph 1	Ph 2	Ph 3	Ph 1	Ph 2	Ph 3
P01	0.6005	0.3361	0.4047	0.5177	0.3634	0.6210	0.5073	0.4268	0.5670	0.7334	0.4812	0.3188
P02	0.6060	0.3594	0.2222	0.8155	0.4687	0.4876	0.4785	0.3931	0.3395	0.4999	0.2313	0.4030
P03	—	0.3642	0.2738	0.7270	0.4333	0.4209	0.7432	0.4201	0.5235	0.3011	0.5566	0.5783
P04	—	—	0.3693	—	—	—	—	—	—	—	—	—
O01	0.4938	0.1850	0.3793	0.3886	0.2415	0.3587	0.6767	0.4869	0.5341	0.3725	0.2313	0.3771
O02	0.7656	0.2514	0.2109	0.8005	0.2698	0.3494	0.6087	0.4294	0.4895	0.7281	0.1501	0.0632*
O03	0.7813	0.4153	0.3645	—	—	0.3419	0.6925	0.4304	0.5181	—	—	—
C01	0.1739	0.1742	0.2540	0.2103	0.2965	0.4965	0.5126	0.4021	0.3147	0.4838	0.2250	0.1797
C02	0.3978	0.1416	-0.023†	0.0673*	0.1802	0.3158	0.4397	0.4679	0.3866	0.3694	0.3917	0.2305
C03	0.4555	0.1513	-0.029†	0.7410	0.3206	0.2811	0.5840	0.4627	0.3022	0.4583	0.2707	0.0174†
C04	0.6867	—	-0.016†	—	—	—	—	—	—	—	—	—

The results indicate that parallel (P) motions maintain the highest average structural fidelity (0.4693 ± 0.1429), suggesting it is the most similar to the original phrases. Oblique (O) motions follow closely with high correlation but greater variability (0.4318 ± 0.1918), while contrary (C)

motions consistently exhibit the lowest correlation (0.3099 ± 0.1835), signifying substantial structural deviation.

Statistical analysis via ANOVA confirmed significant differences in PCC values across motion types ($F = 8.2860$, $p = 0.0005$). Subsequent Tukey HSD tests revealed that while P and O motions are not significantly different from each other structurally ($p = 0.6609$), C motions are substantially different from both (mean difference of 0.1594 and 0.1219 and significance of 0.0005 and $p = 0.0139$, respectively), confirming them as the least structurally faithful to the original material. Interestingly, P motions based on strict transposition often exhibited lower correlation than those based on interval variations, likely due to the linear shifts across the latent dimensions.

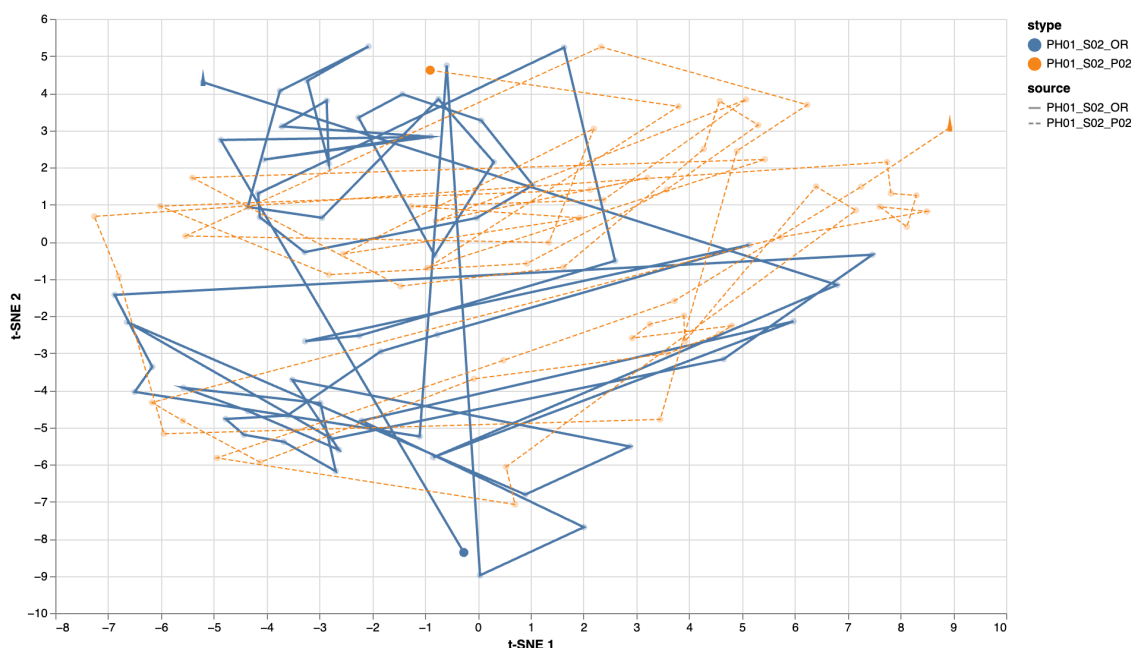


Figure 11.14: Visualization of high-dimensional trajectories via t-SNE, comparing an original phrase with its most structurally similar parallel variation ($PCC = 0.8155$). The grouping illustrates how non-transpositional parallel elaborations maintain cohesive patterns in the latent space. Frame windowing was increased to 12 frames to reduce noise in the two-dimensional representation.

Performance-specific nuances significantly influenced these correlations across the dataset. Saxophonist 02's second parallel motion for phrase 1 achieved the dataset's highest PCC of 0.8155, representing a version characterized by variations in intervals rather than strict transposition. As visualized in the t-SNE projection in Figure 11.14, this sample demonstrates structural similarities in how the latent points traverse space, even though the projection appears rotated. Interestingly, P motions based on strict transposition often resulted in lower structural correlations than those featuring more fluid interval variations, as seen in samples such as PH03_S01_P02 and PH02_S04_P0211. While Saxophonist 03 demonstrated the highest overall correlations across all motion types, as illustrated in Figure 11.15, Saxophonist 01 showed significant variability and lower average values in C motions. This considerable variability in C motions was particularly evident in phrase 3, where Saxophonist 01 mostly presented negative correlations, indicating a

Angular distance provides a scale-independent measure of directional difference, allowing us to evaluate how far a variation deviates from the average intent of the original phrase. As shown in Table 11.3, C motions exhibit the highest level of directional deviation (0.8267 ± 0.4138), followed by P (0.6985 ± 0.3729) and O (0.5780 ± 0.3901) motions.

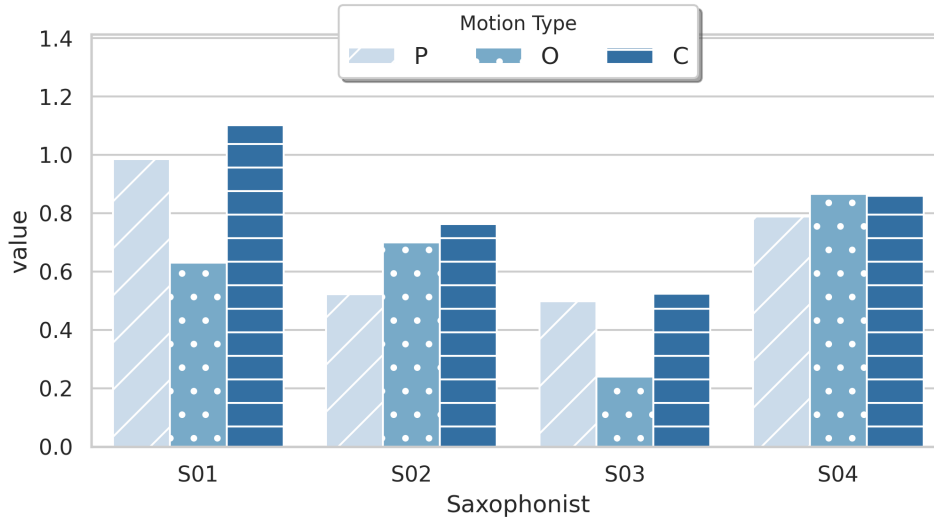


Figure 11.16: Angular distance values between centroids for different motion types (P, O, C) across four saxophonists are analyzed. Lower angular distances indicate a stronger similarity between the recorded motion’s directionalities and the original musical phrase.

While P motions demonstrate some moderate inconsistency in their distances, C motions show a pronounced divergence in direction. Interestingly, for Saxophonists 01 and 03, O motions were more directionally aligned with the original than P motions, a trend further visualized in Figure 11.16. Statistical analysis via ANOVA ($F = 3.3371, p = 0.0395$) confirmed a significant difference in angular distances, with a Tukey HSD test specifically identifying the deviation between C and O motions as statistically significant ($p = 0.0305$). These results suggest that while P and O motions tend to gravitate toward a shared directional path, contrary motions actively break away from this central trajectory, establishing a distinct directional signature.

To further quantify this alignment, we used cosine similarity to measure the angle between vectors across different variations. As detailed in Table 11.4, P motions demonstrated the strongest directional alignment with the original phrases (0.4596 ± 0.1453). In contrast, contrary motions consistently scored the lowest (0.3107 ± 0.1766). In complex sections, such as Phrase 3, the similarity for contrary motions frequently dropped toward zero or even reached negative values, indicating a near-complete reversal of the latent trajectory.

The individual differences among the performers are striking, as illustrated in Figure 11.17. Saxophonists 02 and 04 demonstrate an apparent hierarchical similarity ($P > O > C$), whereas Saxophonists 01 and 03 actually display higher similarity values for O motions than P motions. Notably, Saxophonist 03’s phrases align most consistently with the original across all motion types. ANOVA results ($F = 7.7645, p = 0.0007$) and Tukey’s HSD test confirm that C motions

Table 11.4: Cosine similarity between synchronized latent trajectories of recorded motion types and original phrases. Values closer to 1 indicate stronger directional alignment, whereas values near zero or negative indicate structural divergence.

	Saxophonist 01			Saxophonist 02			Saxophonist03			Saxophonist 04		
	Ph1	Ph2	Ph3	Ph1	Ph2	Ph3	Ph1	Ph2	Ph3	Ph1	Ph2	Ph3
P01	0.6042	0.3705	0.3923	0.4889	0.3546	0.6123	0.5150	0.4323	0.5673	0.7405	0.4580	0.3015
P02	0.6097	0.3919	0.2418	0.8122	0.4377	0.4301	0.4804	0.3981	0.3378	0.5003	0.1971	0.3227
P03	—	0.3737	0.2454	0.7233	0.4004	0.4184	0.7408	0.4123	0.5346	0.2984	0.5181	0.5376
O01	0.5017	0.2426	0.3554	0.3787	0.2353	0.3498	0.6713	0.4753	0.5466	0.3711	0.2678	0.3783
O02	0.7718	0.3054	0.2213	0.7938	0.2839	0.3009	0.6090	0.4235	0.4709	0.7326	0.1601	0.0598
O03	0.7805	0.4358	0.3488	—	—	0.3136	0.6787	0.4464	0.4799	—	—	—
C01	0.1779	0.2237	0.2571	0.2061	0.2862	0.4216	0.5038	0.4090	0.3243	0.4802	0.2623	0.1823
C02	0.4024	0.1614	-0.015	0.0770	0.1489	0.2986	0.4475	0.4467	0.4113	0.3640	0.3856	0.2395
C03	0.4504	0.1842	-0.027	0.7425	0.3408	0.2362	0.5608	0.4265	0.3261	0.4592	0.2880	0.0330
C04	0.6881	—	-0.004	—	—	—	—	—	—	—	—	—

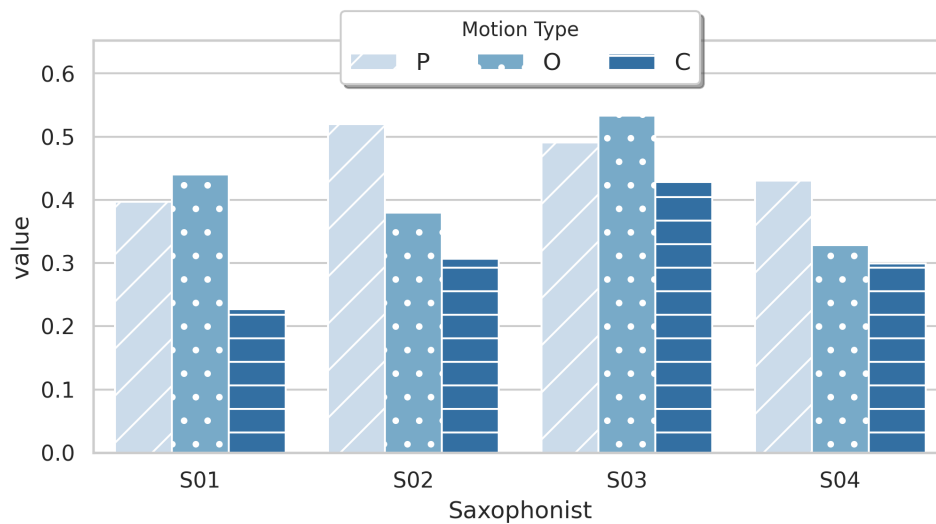


Figure 11.17: Cosine similarity values for different motion types (P, O, C) across four saxophonists are analyzed. Higher cosine similarity indicates a stronger similarity between the recorded motion's directionalities and the original musical phrase.

are significantly different from both P ($p = 0.001$) and O ($p = 0.0305$) motions. The lack of a significant difference between O and P motions ($p = 0.7907$) suggests that while these two types preserve a shared directional trend, contrary motion represents a fundamental redirection of the musical gesture.

These metrics reinforce the structural correlation data: P and O motions generally maintain a common directional trend, whereas C motions involve a radical redirection of the musical gesture within the RAVE manifold.

11.4.2.3 Variation

The final metric used to analyze the Motiv dataset is Euclidean distance, which quantifies the absolute spatial deviation of performative variations from their original anchors. As detailed in Table 11.5, this measure reflects the degree of interpretive stretch within the latent space.

In line with the previous metrics, C motions yielded the highest average spatial variation (3.9112 ± 0.5741), indicating a clear departure from the original trajectory. P and O motions exhibited notably lower and nearly identical average distances (3.5491 ± 0.5179 and 3.5615 ± 0.5857 , respectively), suggesting they remain spatially closer to the original phrase in the latent space, with O motions introducing more structural or directional shifts than P motions.

Table 11.5: Euclidean distances between latent trajectories of recorded motion types and original phrases. Higher values represent a greater absolute spatial deviation (variation) from the original performance manifold.

	Saxophonist 01			Saxophonist 02			Saxophonist 03			Saxophonist 04		
	Ph1	Ph2	Ph3	Ph1	Ph2	Ph3	Ph1	Ph2	Ph3	Ph1	Ph2	Ph3
P01	3.494	3.654	3.763	3.925	4.037	2.712	4.014	3.512	2.817	2.602	3.549	4.029
P02	3.469	3.545	4.121	2.467	3.538	3.267	4.048	3.653	3.505	3.743	4.474	4.271
P03	—	3.609	4.386	2.743	3.694	3.547	2.716	3.422	3.129	4.212	3.073	3.042
P04	—	—	3.988	—	—	—	—	—	—	—	—	—
O01	3.609	4.009	3.987	4.147	4.147	3.615	3.000	3.429	3.062	4.286	4.099	3.690
O02	2.586	3.831	4.085	2.500	3.919	4.016	3.225	3.563	3.248	2.579	4.407	4.817
O03	2.420	3.366	3.584	—	—	3.601	2.926	3.507	3.148	—	—	—
C01	4.703	4.183	4.185	4.396	3.960	3.300	3.557	3.479	3.674	3.168	4.058	4.533
C02	3.828	4.091	5.008	5.023	4.387	3.583	3.719	3.409	3.512	3.494	3.610	4.045
C03	3.488	4.194	5.033	2.660	3.759	3.791	3.671	3.591	3.755	3.487	3.959	4.751
C04	2.779	—	4.800	—	—	—	—	—	—	—	—	—

Consistent with earlier metrics, Saxophonist 01 demonstrated extreme variation in phrase 3, with C motions (C02 and C03) reaching distances above 5.0, as illustrated in Figure 11.18. Conversely, Saxophonist 02 showed much tighter clusters for certain P variations (e.g., P02 in phrase 1), mirroring the high structural correlation found in previous sections. ANOVA results ($F = 4.7884$, $p = 0.0103$) confirmed that the absolute variation significantly differs across motion types, with C motion established as the most spatially divergent category.

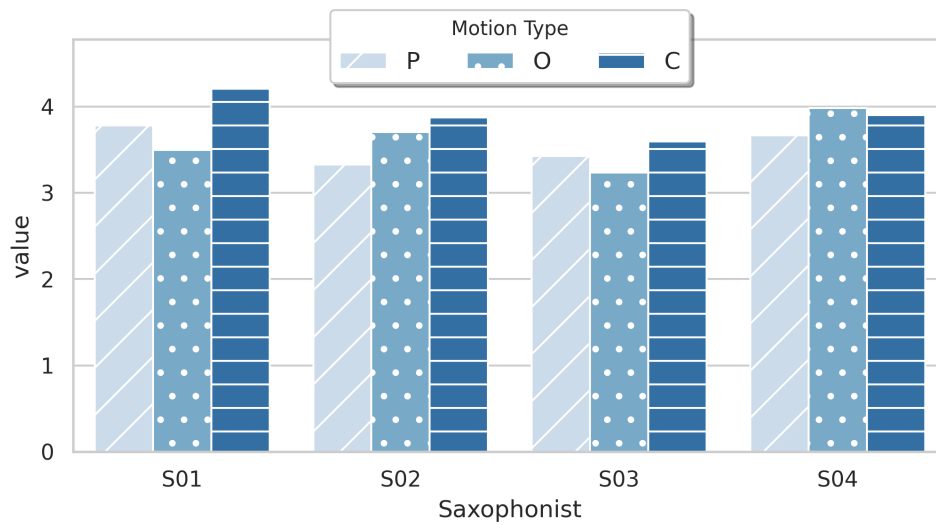


Figure 11.18: Mean Euclidean distances for each player-motion combination. This visualization highlights the variability in deviation from the original phrase across different motion types and performers.

11.5 Summary: Mapping Gestural Intent in the Latent Space

The preliminary analysis of the Motiv dataset (Carvalho et al., 2025a) confirms that musical transformations (parallel, oblique, and contrary) are not merely theoretical constructs but manifest as distinct, statistically significant geometric signatures within the latent space. By utilizing a RAVE model to map recorded audio, we have established a structured representation of musical phrases that enables the objective analysis of thematic development.

11.5.1 Summary of Findings

The exploratory results reveal a clear hierarchy of structural and directional relationships relative to the source material. Parallel motions tend to preserve the strongest structural similarity and sense of directionality, highlighting their role in maintaining the integrity of the original phrase. In contrast, contrary motions exhibit the most significant structural deviations and spatial dilation, representing a radical divergence in both directionality and absolute Euclidean distance. Oblique motions demonstrate a mixed pattern, exhibiting both resemblance and deviation, which indicates a more complex relationship influenced by individual performer interpretations.

Furthermore, the study revealed that individual differences among the four professional saxophonists significantly influenced the execution of these motion types. These variations underscore the importance of considering performer-specific nuances, as musical gestures are deeply tied to a performer's unique expressive style.

11.5.2 Implications for Co-Creative Systems

While a larger dataset would be required to robustly train a model to fully learn underlying motion dynamics, the emergence of statistically significant results indicates that the current volume of data is sufficient for exploratory analysis and real-time mapping. The Motiv dataset provides the necessary topological map for navigating latent spaces, moving beyond simple stochastic sampling toward intentional musical dialogue.

By training generative models on this semantically labeled data, it becomes possible to create systems that synthesize innovative improvisations aligned with specific motion types, allowing for co-creative agents that function as a contextually aware partner, capable of recognizing a performer's trajectory and responding with appropriate variations or contrasts. Ultimately, this research contributes to a deeper understanding of how musical motions can be synthesized and analyzed, pushing the boundaries of AI-driven performance.

Chapter 12

Aethra: A Mutable Interface for Mixed Music Navigation

The culmination of our research in Part III is the development of Aethra, a co-creative system designed to address the performative limitations inherent in traditional mixed music, specifically its reliance on fixed electroacoustic media. As we have reviewed in Chapter 2, conventional mixed music often limits a performer’s interpretive freedom due to the rigid temporal and timbre-nuanced nature of pre-recorded electronics, effectively reducing the musical work to a static artifact and positioning the computer as a passive execution tool.

To overcome these rigid constraints, Aethra is designed to transform the electronic medium from a static backing track into a responsive, co-creative partner. By shifting the computational paradigm from static audio playback to dynamic latent space navigation, the system operates as a mutable interface. This approach allows the electronic architecture to algorithmically adapt to the live musician’s input in real time, establishing an active, co-improvisational environment that restores expressive and stylistic agency to the performer.

The chapter is structured in the following manner. Section 12.1 details the core architecture and operational workflow of the Aethra system, delineating its bifurcated pipeline into an offline latent space preparation phase and an online Pure Data-driven performance environment. This section also introduces the system’s unified one-knob interface and its dual-mode navigation engine, encompassing both distance-based and motion-based algorithmic strategies, alongside its granular playback synthesis mechanism. Section 12.2 presents the performance case study *Echoes of the Unseen*, an AI-mediated co-improvisation for tenor saxophone and live electronics built upon a reinterpretation of Jesús Villa-Rojo’s *Lamento*. Here, the system’s creative affordances are evaluated in a live performative context, examining the real-time modulation of neural timbre transfer archetypes, the use of structural auto-scores to manage improvisational vulnerability, and the integration of visual latent mapping via TouchDesigner to bridge the perceptual gap between the algorithmic processes and the audience. Finally, Section 12.3 synthesizes these technical and artistic outcomes, reflecting on the emergence of the post-performer paradigm. It concludes by

discussing how Aethra redefines the relationship between live musicians and fixed electronic media, transforming static electroacoustic tape into an active, co-creative improvisational partner.

12.1 System Overview and Core Architecture

Aethra is an interactive audiovisual system that integrates machine learning, latent space navigation, and real-time sound synthesis for mixed music performance. Designed to replace static audio playback with dynamic latent space exploration, the electronic architecture algorithmically adapts to the live musician’s acoustic and gestural input in real time, establishing an active, co-improvisational environment.

The system’s core logic relies on integrating the RAVE architecture with Pure Data (PD)¹ via the py4pd² library. The system architecture, as illustrated in Figure 12.1, is divided into two primary operational stages: 1) an offline Google Colab-based preparation phase³ and 2) an online PD real-time performance system for live improvisation and synthesis.⁴

Furthermore, the architecture incorporates a visual engine developed in TouchDesigner, which receives Open Sound Control (OSC) data streams from the PD environment to render a real-time projection of the navigated latent space. This visual component functions as an interface for multimodal explainability, providing a direct bridge between the system’s invisible algorithmic navigation and the audience’s perception.

The subsequent sections detail the technical implementation of these core architectural components. Particular emphasis is placed on the Latent Space Navigator and its single-axis control interface, outlining how the system translates continuous manual input into algorithmic spatial queries within the neural manifold. Demonstration videos are available at <https://nadiacarvalho.github.io/MutableMusicalInterface/#/software/aethra>.

12.1.1 Offline Preparation Phase

The offline preparation phase of the Aethra system automates the extraction of latent representations from a fixed audio tape. This procedure consists of three operational stages, converting high-resolution raw audio waveforms into a serialized, navigable corpus.

The processing pipeline starts by routing a target audio file (i.e., the fixed tape) through the encoder of a pre-trained RAVE model. The raw audio is processed at the model’s native sample rate (e.g., 44.1 or 48 kHz), where the encoder compresses the high-resolution waveform into n -dimensional latent vectors (typically 8 or 16). This process yields discrete latent frames at a rate of

¹<https://msp.ucsd.edu/software.html>, last accessed January 29, 2026.

²<https://charlesneimog.github.io/py4pd/>, last accessed January 29, 2026.

³<https://github.com/NadiaCarvalho/Aethra/blob/main/Aethra-LatentSpaceCreator.ipynb>, last accessed January 29, 2026.

⁴<https://github.com/NadiaCarvalho/Aethra/tree/main/Aethra-PD>, last accessed January 29, 2026.

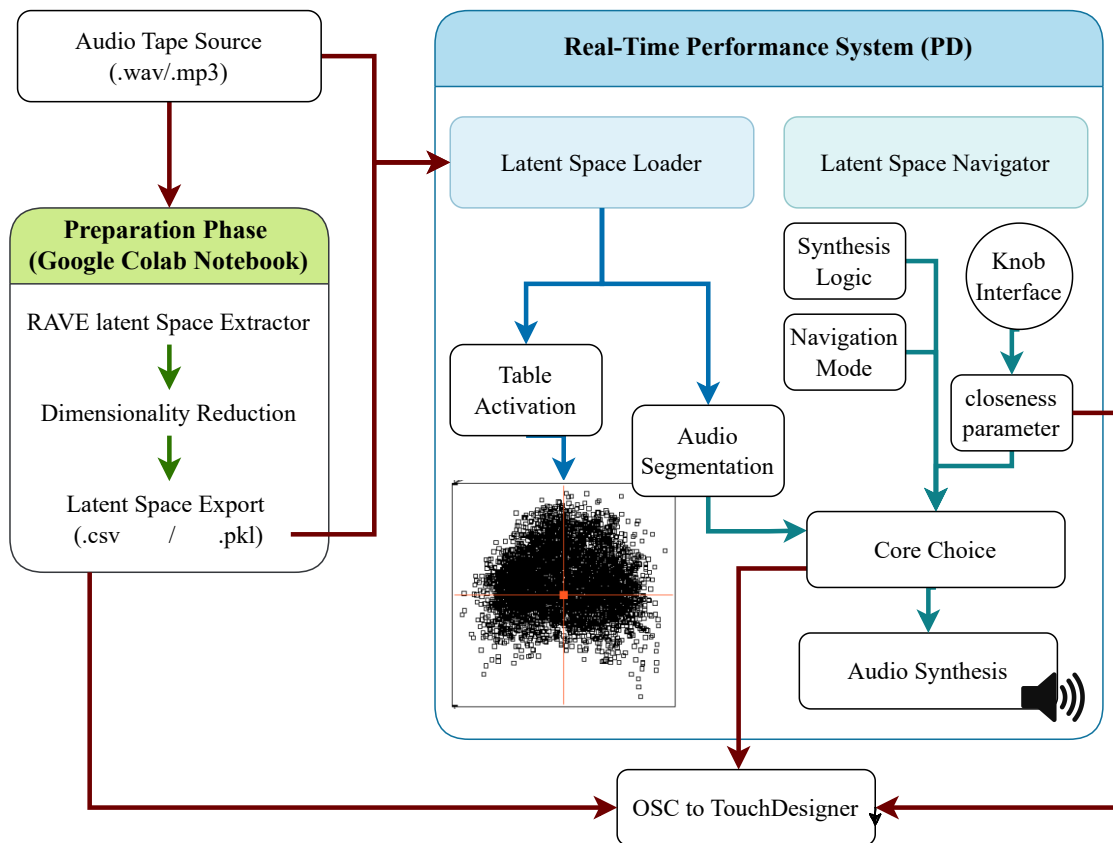


Figure 12.1: Architectural Overview of the Aethra System. The workflow is bifurcated into two distinct operational stages: (left) the offline preparation phase, utilizing a Google Colab notebook to encode the fixed audio tape into serialized latent representations; (right) the online performance phase, operating within PD via the py4pd library. The real-time environment facilitates navigation of the preloaded latent space via a single-axis control knob. Additionally, the architecture incorporates a dual-mode navigation engine, dynamic latent dimension weighting, real-time granular audio synthesis, and an OSC data stream to a TouchDesigner visual engine for real-time topological rendering.

approximately 23.4 frames per second (corresponding to one frame every 43 milliseconds). Subsequently, rather than grouping raw waveform samples, the system calculates the mathematical average of every four consecutive latent frames. This temporal aggregation generates a unified latent coordinate representing approximately 170 milliseconds of audio. Consequently, this architectural procedure decreases the total number of spatial coordinates by a factor of four, significantly optimizing computational efficiency for real-time navigation without compromising morphological continuity.

To facilitate visual navigation and verify the continuous spatial distribution of the learned manifold, topological mapping is conducted utilizing the Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction algorithm. UMAP is employed because, unlike linear models (such as PCA) or local-centric methods (such as t-SNE), it operates under the assumption that data is uniformly distributed across a locally connected Riemannian manifold. This mathemat-

ical framework balances the preservation of local neighborhood similarities, ensuring that related timbre motifs remain clustered, with the maintenance of global topological distances within the non-linear continuous space.

By projecting these high-dimensional continuous vectors into a 2D coordinate system, the algorithm reveals dense regions of similar timbre motifs, enabling visual analysis of the material's structural density without imposing artificial categorizations. This 2D reduction serves as the primary data input for the TouchDesigner visual engine. The reduced coordinates are exported as a serialized dataset and routed into the visual interface via a Table DAT to drive Geometry Instancing, establishing a persistent visual map of the complete timbre corpus. Ultimately, this integration functions as a transparency layer, providing a navigable projection of the system's algorithmic processing for both the performer and the audience.

The preparation phase concludes with the exportation of the corpus into serialized Pickle (.pkl) and CSV (.csv) formats, each fulfilling a specific architectural function in bridging the offline and online stages. The .pkl format is explicitly designated for the Aethra-PD system, as it skips the extra parsing steps of text-based formats, letting complex data load straight into memory. This optimization enables the py4pd objects to execute high-speed spatial searches efficiently during real-time performance. Conversely, the CSV (.csv) file is generated specifically for TouchDesigner, supplying the dimensionally reduced 2D spatial coordinates required to drive the visual engine. By offloading the computationally intensive processes of neural feature extraction and dimensionality reduction to the offline Google Colab environment, the real-time architecture is optimized for performance. Consequently, this allows the Latent Space Loader in PD to dedicate computational resources entirely to low-latency navigation and audio synthesis during live execution.

12.1.2 Online Performance System

Following the offline preparation phase, the online performance phase integrates the static latent map into a real-time, interactive environment within PD. This integration is handled by the Aethra-PD architecture, whose primary control interface is illustrated in Figure 12.2. Aethra-PD leverages the py4pd library to couple the deterministic audio synthesis engine of PD with the high-level data processing capabilities of Python.

Before detailing the specific technical modules, a high-level overview of the Aethra-PD architecture is provided. As illustrated in the online performance system panel of Figure 12.1, the system functions as a continuous computational pipeline that translates physical control gestures into synthesized audio. The primary objective of this architecture is to abstract the multidimensional complexity of the neural network, reducing cognitive load and facilitating direct interactive navigation.

The workflow initiates with the Latent Space Loader, which imports the serialized offline data to manage table activation and audio segmentation. Following initialization, the performer utilizes the single-axis control interface to modulate the *closeness* parameter. This control data is routed to the Latent Space Navigator, which, governed by the active navigation mode and synthesis logic,

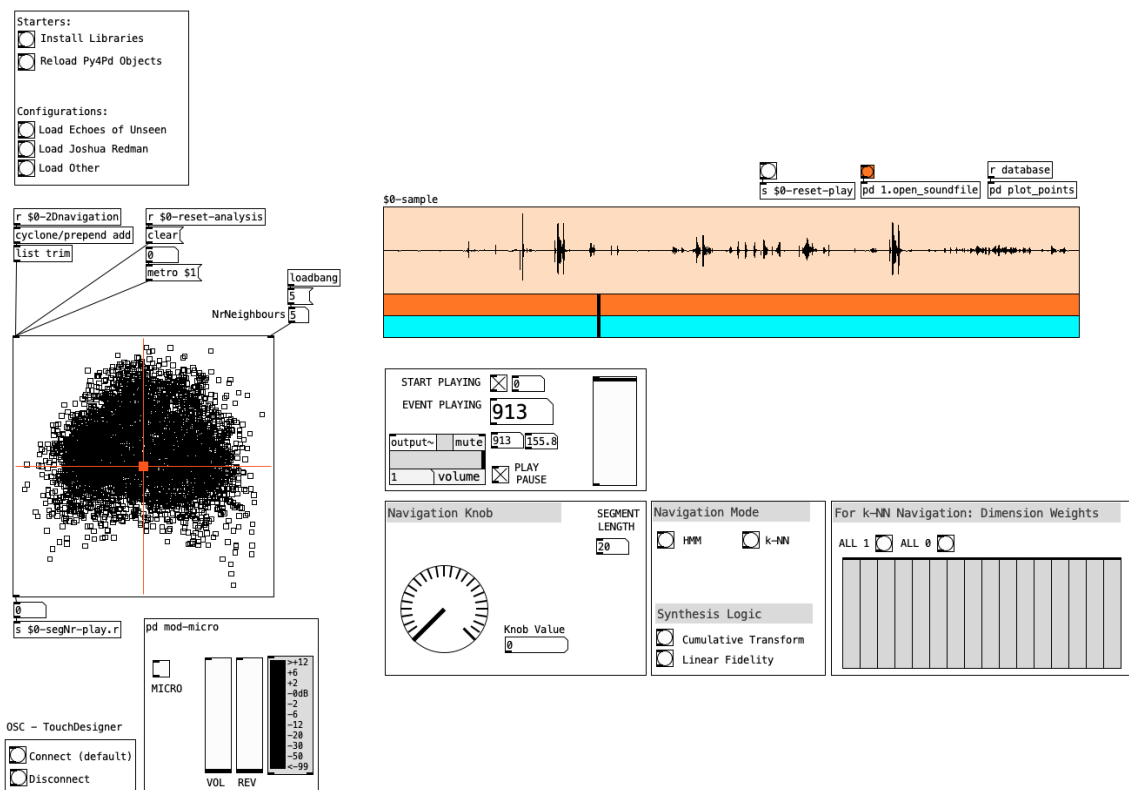


Figure 12.2: Main Performance Interface of the Aethra System in PD. It centralizes control of the navigation in a single knob interface. The user can also toggle between the two different navigation strategies (i.e., HMM/ k -NN), between synthesis logic modes (i.e., linear fidelity/cumulative transformation), and set latent dimension weights for proximity calculation.

computes the target latent coordinate to be sampled. Finally, this coordinate is transmitted to the audio synthesis engine to generate the acoustic output.

The initialization of the live environment is managed by the Latent Space Loader (via the py4pd object `py.load_latent`), which serves as the primary data interface between the serialized offline data and the PD audio synthesis engine. This object performs three critical tasks to ensure the system is ready for performance. First, it parses the serialized .pkl file produced during the offline Google Colab preparation phase, reconstructing the n -dimensional latent vectors and synchronizing them with their original temporal metadata to establish the continuous navigable manifold.

To streamline the initialization of this computational pipeline, the Latent Space Loader utilizes a configuration file that specifies the paths for the core assets required for real-time execution, as shown in Figure 12.3. This file automates the initialization procedure by referencing the source audio tape, the serialized latent vectors (.pkl files), and the HMM model trained on the Motiv dataset (.pkl) for temporal modeling. By centralizing these directory paths, the architecture establishes a decoupled framework, enabling the integration of distinct musical corpora or pre-trained models into the system without necessitating the manual reconfiguration of the underlying PD objects or Python scripts.

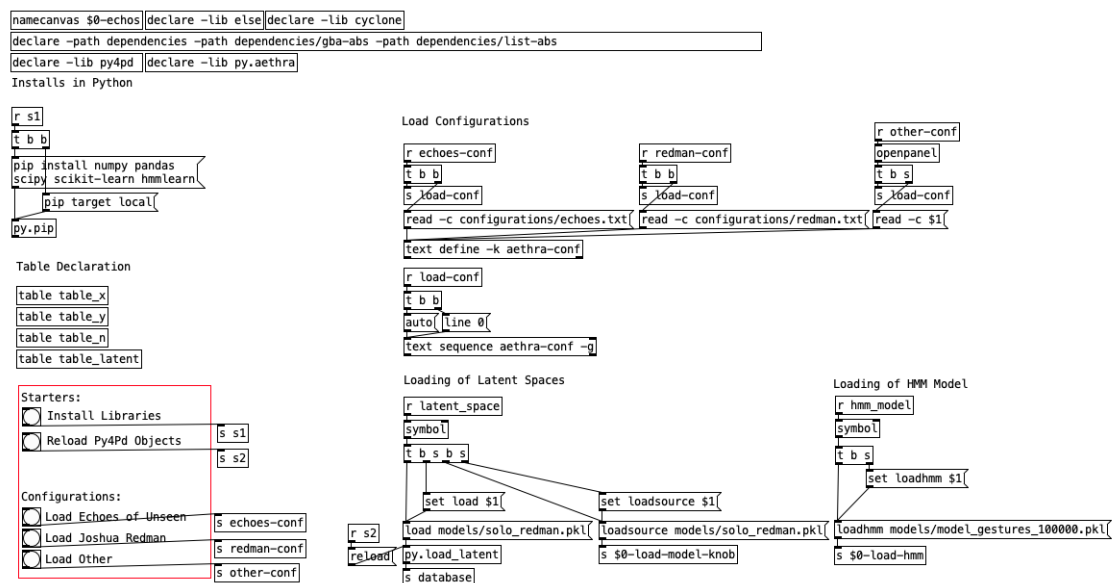


Figure 12.3: PD Interface for Asset Initialization and Global Configuration. The interface displays the management of core system dependencies, including the Latent Space Loader and the audio tables. The initialization relies on a configuration parser that automates the retrieval of file paths for the audio source, saved latent vectors (.pkl files), and HMM transition models. This way, everything can work smoothly together between the Python backend and the real-time synthesis engine.

As a second step, the Latent Space Loader populates native PD tables (i.e., `table_x`, `table_y`, `table_n`, and `table_latent`) to establish the performative interface. `table_x` and `table_y` are populated with the reduced 2D spatial coordinates (used to render the visual map in the PD interface), while `table_n` stores the corresponding audio unit indices. Crucially, `table_latent` is utilized to store the real-time latent dimension weights as an array of dynamic floating-point values that the performer adjusts via a multi-slider graphical user interface (GUI) to prioritize specific acoustic features during distance calculations. Because different RAVE models possess varying numbers of latent dimensions (e.g., 16 for VocalSet, 8 for Strengjaver), the array size of `table_latent` is automatically resized during initialization to match the specific dimensionality of the loaded model. This process generates a functional interface for the performer, transforming the abstract high-dimensional manifold into a visual representation of timbre clusters. By mapping these neural representations to a tangible GUI, the system provides a clickable, navigable map within the PD patch that allows for intuitive gestural control over the latent space. At last, complex data is controlled in the background by a persistent `py4pd` object. This setup ensures low-latency performance, even during intense improvisations, as it enables us to rapidly access complex data without slowing down PD's real-time audio processing.

Following data initialization, the Latent Space Navigator (`py.navigation_mechanism`) directs the synthesis pipeline to compute the target latent coordinate. To address the phenomenological challenges associated with navigating complex, high-dimensional generative spaces, the architecture features a single-axis control interface. By compressing the n -dimensional RAVE

manifold into this unified control axis, the system mitigates hermeneutic overload and reduces the performer’s cognitive burden. This design establishes a balance between embodiment and hermeneutic relations, enabling the performer to intuitively gauge their topological distance from the source material and modulate expressive trajectories without losing their physical connection to the performative act.

This control framework is governed by a dual-mode engine integrating the navigation mode (distance-based vs. motion-based strategies) and the synthesis logic (linear fidelity vs. cumulative transformation), which builds upon the navigational strategies evaluated in Chapters 10 and 11 and summarized in Table 12.1. By integrating these control streams, the system translates the high-dimensional mathematical constraints of the latent space into a navigable, real-time performative environment.

Table 12.1: Comparison of Navigation Strategies within Aethra-PD.

Feature	Distance-Based Strategy	Motion-Based Strategy
Primary Algorithm	Weighted k -nearest neighbors	Gaussian HMM (Full Covariance)
Control Value C	Continuous (Exploratory Drift)	Discrete (Motion Selection)
Latent Mapping	Weighted Euclidean Distance	Probabilistic State Transitions
Temporal Logic	Static / Spectral Focus	Predictive / Phrase-based
Musical Application	Textural Sculpting	Structural / Gestural Variation

12.1.2.1 Distance-Based Navigation Strategy

The distance-based strategy functions as the primary computational paradigm for continuous latent navigation. In this operational mode, the system processes the manifold as a continuous parameter space, wherein the performer’s input dictates the degree of topological deviation from the source material. A fundamental characteristic of this architecture is the capacity to dynamically adjust latent dimension weights. By modulating these weights, the performer can prioritize specific acoustic features, such as pitch stability or spectral brightness, during the weighted Euclidean distance calculation, as detailed in Algorithm 1, enabling selective variation across the remaining dimensions.

Drawing upon the distance-based navigation strategies evaluated in Chapter 10, the system segments the sorted latent data into twenty-five discrete bins, as detailed in Algorithm 1. The continuous control value functions as an index selector to designate a specific bin corresponding to the targeted spatial neighborhood. Within this designated bin, the algorithm executes a weighted random selection, assigning higher probability weights to coordinates with the minimal Euclidean distance from the reference target. Ultimately, this computational strategy facilitates controlled topological variation and exploration while dynamically adhering to the performer’s structural and expressive constraints.

Algorithm 1 Distance-Based Navigation Strategy Pseudo-Code.

Require: L (Latent Dimension Weights), D (Pre-computed Distance Matrix), C (Closeness Control $[0, 1]$)

Ensure: s (Selected Unit ID)

```

1: {1. Calculate Weighted Euclidean distances for all units}
2: for each unit  $j$  in dataset do
3:    $dist[j] \leftarrow \sqrt{\sum_{i=1}^{16} L_i \cdot D_{i,j}}$ 
4: end for
5: {2. Sort and Segment Manifold}
6:  $ids \leftarrow$  sort indices of  $dist$  in ascending order
7:  $N_{bins} \leftarrow 25$ 
8:  $size_{bin} \leftarrow \text{length}(ids) / N_{bins}$ 
9: {3. Map User Input to Bin Index}
10:  $idx_{bin} \leftarrow \text{floor}(\text{clamp}(C \cdot (N_{bins} - 1), 0, N_{bins} - 1))$ 
11: {4. Inverse-Distance Weighted Selection}
12:  $S \leftarrow ids[idx_{bin} \cdot size_{bin} \dots (idx_{bin} + 1) \cdot size_{bin}]$  {Target segment}
13:  $d_{max} \leftarrow \max(dist[S])$ 
14: for each  $id$  in  $S$  do
15:    $w_{id} \leftarrow (d_{max} - dist[id]) + 10^{-6}$  {Inverse weighting}
16: end for
17:  $s \leftarrow \text{weighted\_random\_choice}(S, weights = w)$ 
18: return  $s$ 

```

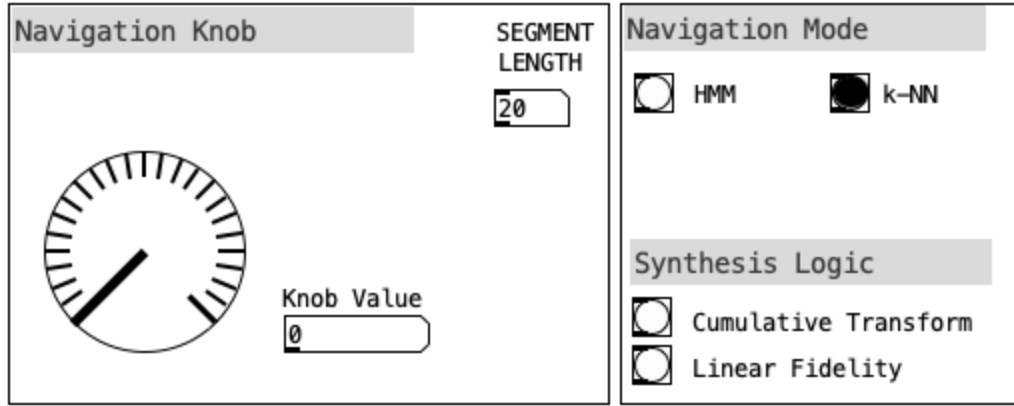


Figure 12.4: One-knob interface C during distance-based exploration. The ticks in the knob demonstrate that C works as a continuous value, mapping the input to latent similarity bins.

12.1.2.2 Motion-Based Navigation Strategy

In the motion-based navigation strategy, the architecture utilizes a Gaussian HMM trained on the Motiv dataset. The integration of an HMM addresses the inherent limitations of purely proximity-based sampling algorithms, such as standard k -NN or distance-based binning. Because these proximity strategies navigate the latent space through isolated, step-by-step statistical matching, they fundamentally lack the compositional foresight and anticipatory behavior required for coherent structural development.

We selected a Gaussian HMM because it is able to map the continuous, high-dimensional RAVE latent space into a structured temporal grammar. Unlike discrete models, the Gaussian HMM utilizes continuous emission distributions to sample latent vectors (Rabiner, 1989). At the same time, its full covariance matrix preserves the intricate statistical dependencies between timbre dimensions (e.g., correlation between spectral brightness and harmonic density), ensuring that synthesized variations remain within the physical and expressive limits of the source material (Bilmes, 1998). By encoding the transition likelihoods of the archetypal musical motions detailed in Section 11.1, the HMM models the gestural logic of the musical work as a continuously evolving sequence rather than a collection of disconnected acoustic frames. This architectural capability allows the system to compute stylistically coherent trajectories in real time, effectively functioning as an anticipatory mechanism that preserves the intended gestural morphology even during significant structural deviations.

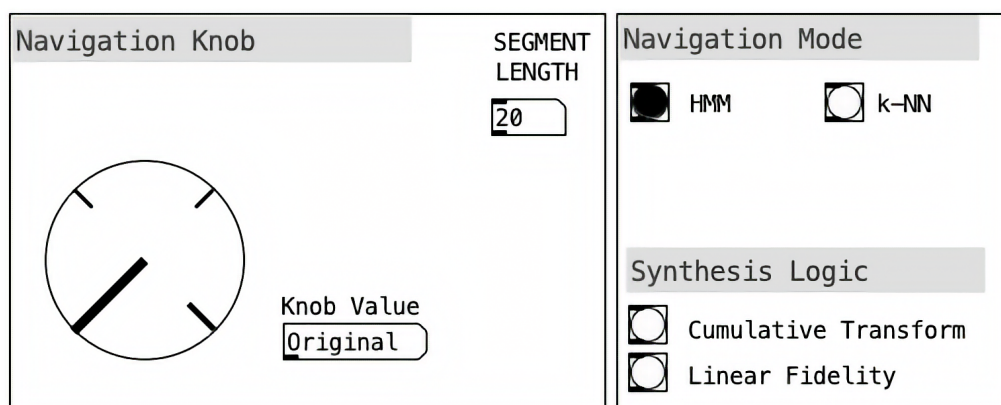


Figure 12.5: One-knob interface *C* during motion-based Mode. The ticks in the knob show how it acts as a discrete selector for original, parallel, oblique, and contrary HMM models.

To construct the HMM model, the n -dimensional latent sequences are organized to train a distinct model for each taxonomic motion type. The architecture employs a full covariance matrix for the multivariate latent sampling process. Unlike a diagonal covariance matrix (Σ), which assumes statistical independence among latent features, this full covariance matrix ($\Sigma \leftarrow H.covars[q_{t+1}]$) explicitly models the correlations between dimensions, such as the relationship between spectral brightness and harmonic density. As detailed in Algorithm 2, this mathematical formulation ensures that the predicted latent vectors remain strictly within the realistic manifold of the instrument’s expressive range. Consequently, the system effectively captures complex, non-linear dependencies across multiple timbre dimensions.

In a performance setting, the control value *C* functions as a discrete selector to alternate between the specific pre-trained motion models. Upon the selection of a motion trajectory, the system decodes the most probable hidden state sequence. It then computes the subsequent latent observation by sampling from a multivariate normal distribution, which is parameterized by the model’s learned mean vectors and full covariance matrices. Because the HMM generates a continuous synthetic latent vector that may not correspond to an exact empirical coordinate in the training

Algorithm 2 State-Space Transition and Observation Sampling Pseudo-Code.**Require:** M (HMM model set), X_{init} (Initial sequence), S_{source} (Search space)**Ensure:** idx (Selected unit index)

```

1: {1. Model Retrieval and State Decoding}
2:  $H \leftarrow \text{SelectModel}(M, \text{motion})$ 
3:  $\Gamma \leftarrow \text{ViterbiDecode}(H, X_{init})$  {Recover hidden state sequence}
4:  $q_t \leftarrow \Gamma[-1]$  {Identify most recent hidden state}
5: {2. Predict Next State}
6:  $P_{trans} \leftarrow H.transmat[q_t]$  {Transition probabilities for current state}
7:  $q_{t+1} \leftarrow \text{RandomChoice}(\text{states}, \text{probabilities} = P_{trans})$ 
8: {3. Multivariate Latent Sampling}
9:  $\mu \leftarrow H.means[q_{t+1}]$  {16-dimensional mean vector}
10:  $\Sigma \leftarrow H.covars[q_{t+1}]$  {Full covariance matrix}
11:  $\epsilon \leftarrow \mathcal{N}(\mathbf{0}, \Sigma)$  {Sample multivariate noise}
12:  $x_{next} \leftarrow \mu + \epsilon$ 
13: {4. Manifold Projection (Nearest Neighbor)}
14:  $idx \leftarrow \underset{j}{\operatorname{argmin}} (\text{EuclideanDist}(x_{next}, S_{source}[j]))$ 
15: return  $idx$ 

```

data, the architecture executes a post-prediction nearest-neighbor search, defined mathematically as a manifold projection, to retrieve the closest matching audio unit from the source corpus. This architectural design guarantees morphological consistency, even during significant structural deviations from the fixed electroacoustic tape. Ultimately, it preserves the stylistic integrity of the musical work while enabling the computational system to exhibit anticipatory behavior and compute contextually coherent trajectories in real time.

12.1.2.3 Synthesis Logic: Linear vs. Cumulative

Both navigation strategies, as illustrated in Figure 12.6, operate through a dual-mode synthesis logic that governs how the anchor coordinate for the subsequent audio unit is calculated. In linear fidelity mode, the system preserves a strict temporal reference to the original score. The subsequent audio unit is retrieved according to its chronological index within the original electroacoustic tape, relative to the global playback position. Consequently, even if a preceding step executes a discontinuous topological deviation into a distant region of the manifold, the system immediately realigns with the deterministic timeline for the subsequent calculation.

Conversely, the cumulative transformation mode employs a recursive computational approach: it evaluates the trajectory of the next audio unit relative to the immediately preceding state, rather than its global chronological index. This architectural design enables the synthesized audio to generate autonomous, continuous morphological evolutions while rigorously maintaining stylistic and topological coherence.

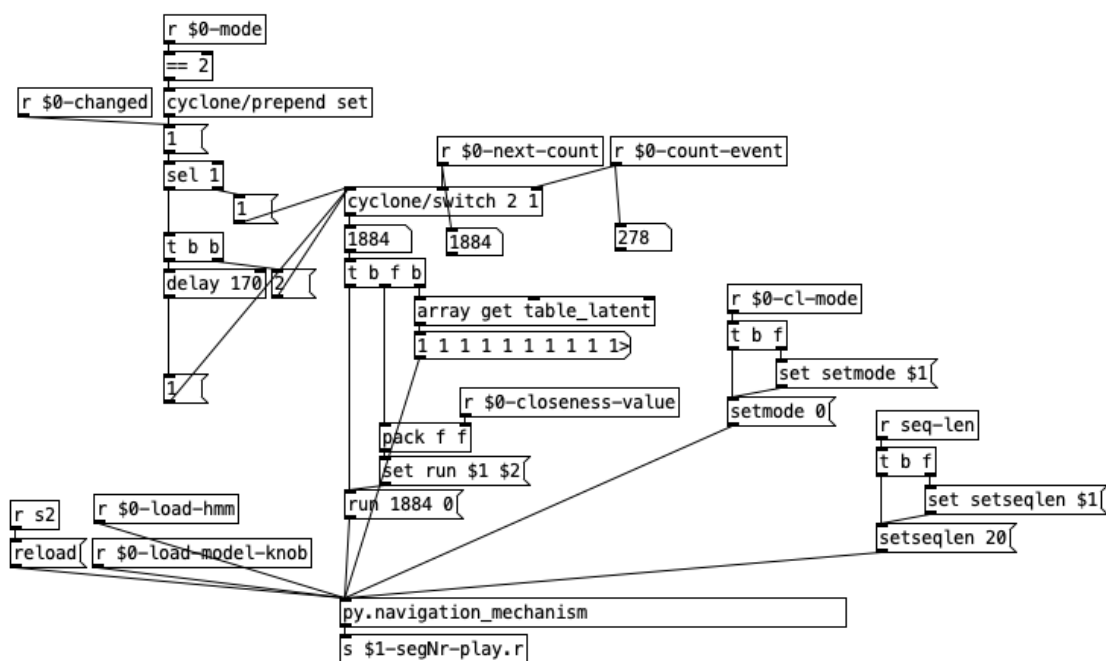


Figure 12.6: Architecture of the navigation logic and synthesis control. The patch integrates the `py.navigation_mechanism` to facilitate real-time latent space selection. The system distinguishes between navigation modes, motion-based and distance-based (`$0-cl-mode`), and synthesis logic, where `$0-mode` toggles between Cumulative (generative sequencing) and Linear (score-following) synthesis logic.

12.1.2.4 Temporal Resolution and Segment Length

The operational dynamics of the navigation engine are governed by the `segment length` parameter. It dictates the temporal duration the system maintains a specific generative trajectory before computing a new spatial target. By processing the audio in discrete macroscopic segments rather than updating the latent coordinate at every micro-unit, the architecture mitigates the morphological discontinuity and acoustic artifacts associated with rapid granular fluctuations.

This parameter directly modulates the system’s interactive latency: shorter segments yield higher reactivity to the performer’s input, whereas longer segments facilitate structural development, allowing a synthesized phrase to achieve morphological completion before a trajectory shift. By calibrating the segment length to the underlying rhythmic or metrical structure, the composer ensures that algorithmic transitions align with significant musical boundaries. Ultimately, this synchronization integrates the complex computational logic with the embodied performance, ensuring the synthesized output remains structurally coherent and contextually bound to the acoustic environment.

12.1.2.5 Granular Playback Engine: From Latent Selection to Sonic Synthesis

The sonic realization of the navigated latent space is executed by a high-performance granular synthesis engine. Representing a deliberate architectural departure from the standard RAVE pipeline,

the system bypasses the neural decoder to function as a CBCS environment. Rather than synthesizing new waveforms from scratch, which can occasionally introduce generative artifacts or spectral blurring, the engine retrieves and concatenates the raw audio recordings associated with the selected latent indices. This design ensures that the system maintains the absolute acoustic fidelity, morphological richness, and authentic expressive nuances of the source material.

In practice, the engine stores the source material in a primary audio buffer (`$0-sample`), accessing specific audio units based on a segment index (`segNr`) provided by the Python navigation layer. By multiplying this index by a constant unit size (8192 samples), the system calculates the precise read-position within the buffer (see Figure 12.7). The unit is retrieved and played with a cloned granular playback abstraction (`abs/player.n`). A Hann windowing strategy (`$0-hann`) is employed to ensure morphological continuity during real-time navigation of the latent space (see Figure 12.7). This windowing approach effectively eliminates spectral clicks during rapid, non-contiguous topological jumps.

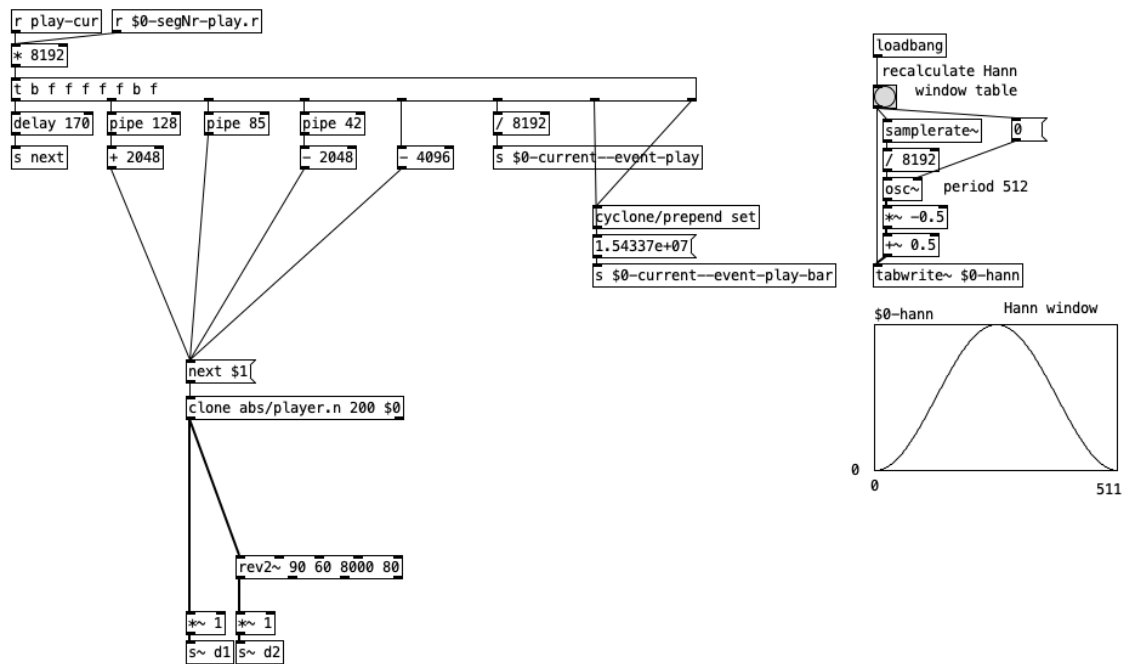


Figure 12.7: Segment indexing and buffer addressing logic. The integer segment index (`segNr – play.r`) received navigation layer is multiplied by a constant unit size of 8192 samples to determine the precise start-read position within the `$0-sample` audio buffer. The unit is retrieved and played with a cloned granular playback abstraction (`abs/player.n`). A Hann windowing strategy (`$0-hann`) is employed to ensure morphological continuity during real-time navigation of the latent space.

A master metronome, defaulting to a 170ms cycle, drives the temporal density of these granular events, creating a smooth and evolving sound (see Figure 12.8). Regular triggering helps blend different segments from the latent space into a unified sound while keeping a consistent shape as the synthesis events flow continuously. Dynamic reverberation modules further facilitate the real-time spatialization of the resulting acoustic textures.

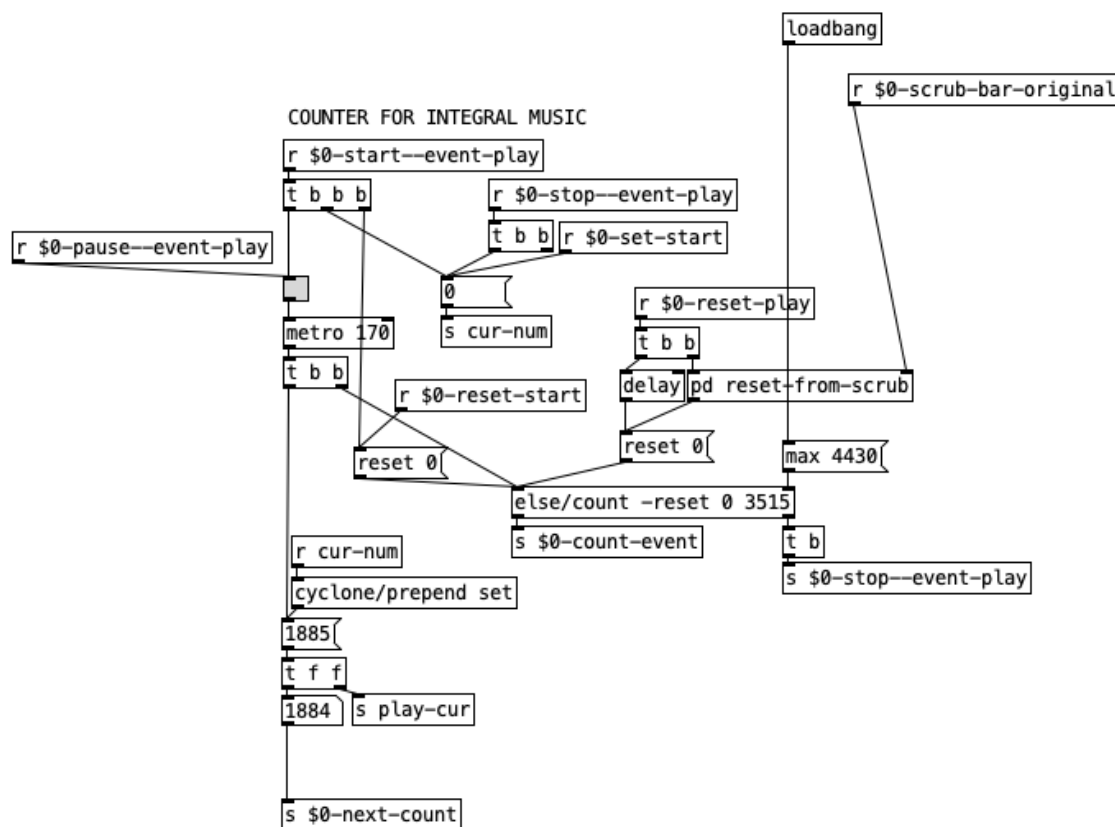


Figure 12.8: Schematic of the real-time granular playback engine. The 170ms metro cycle controls how often the granular envelopes activate, creating a smooth and evolving sound. This regular triggering helps blend different segments from the latent space into a unified sound while keeping a consistent shape as the synthesis events flow continuously.

Following the granular synthesis stage, the post-processing phase routes the generated stereo units through a primary reverberation module, which functions as an acoustic binding agent for the discrete granular textures. In live performance contexts, the architecture accommodates an additional reverberation unit (see mod-micro module in Figure 12.2). By dynamically modulating the dry/wet ratio of this secondary module in correlation with the continuous closeness parameter, the performer can actively manipulate the perceived spatiality of the computational system, transitioning fluidly from a dry, intimate acoustic shadow to a vast, diffuse environmental soundscape. Ultimately, this comprehensive granular architecture transforms the static latent corpus into a highly plastic, living sound-body, providing a robust operational framework for the performer to navigate complex timbre manifolds with immediate, embodied acoustic feedback.

12.1.2.6 Aethra's Decision Mechanism: Arbitrating Latent Navigation

The decision mechanism is primarily a performer-triggered architecture, placing the responsibility of structural and strategic shifts directly in the hands of the musician. By directing the system's latent space navigation, this mechanism effectively functions as the high-level control layer

for an AI-informed CBCS environment. Unlike traditional CBCS frameworks, which organize audio corpora using manually defined acoustic descriptors and frequently navigate these multidimensional spaces blindly through step-by-step statistical matching, Aethra leverages a continuous neural manifold to drive the concatenative process.

Rather than delegating the transition between navigation paradigms (distance/motion-based) to an autonomous algorithm, the system relies on the performer to act as the high-level supervisor. When the performer modulates the control interface, the decision mechanism dictates which latent coordinates are targeted. Aethra explicitly adopts this CBCS framework by bypassing the typical neural architecture decoder. Instead of synthesizing new waveforms from scratch, the system utilizes a granular engine to retrieve and concatenate the actual raw audio recordings associated with the selected latent indices.

This design ensures that the machine functions as an intuitive extension of the performer's creativity. A musician can initiate performance by employing a distance-based technique to retrieve acoustically proximal corpus grains, and subsequently rapidly alter the concatenated output by switching the navigation mode to explore trajectory-aware motions. A background interpolation buffer facilitates this manual arbitration by smoothing the transition when the performer initiates a change. Even though the decision is manual, the system handles the underlying mathematics to ensure that the new strategy inherits the latent coordinates of the previous state, preventing jarring morphological shifts in the concatenated timbre. Consequently, the performer can treat the distance-based and motion-based strategies as distinct operational modes of the same CBCS instrument, facilitating a fluid movement between micro-timbre exploration and macro-gestural development.

Simultaneously, the system facilitates a hermeneutic relation by providing a legible readout of the machine's autonomous logic. The interface functions as an integral component of Explainable AI (XAI), offering the performer a music-focused projection of the model's structural targets. By transforming the abstract calculations into a clear visual and spatial map, the performer is empowered to anticipate the machine's algorithmic suggestions and respond with precise, contextually aware gestures. Ultimately, this balance prevents the performer from feeling disconnected from the system's contextually bounded possibilities, allowing the mutable work to emerge through a continuous, intelligible dialogue between human and machine actants.

12.2 Case Study: *Echoes of the Unseen*

The theoretical and technical foundations of Aethra encounter their practical application in the case study of *Echoes of the Unseen*. *Echoes of the Unseen* is a performance that explores the boundaries of AI-driven improvisational music. Conceived and performed by Jorge Sousa (tenor saxophone) and Nádia Carvalho (live electronics), this work serves as a live demonstration of how timbre latent spaces can be navigated to create a symbiotic relationship between acoustic and digital expression.

12.2.1 Material Preparation

A central element of this case study is the relationship between the live instrument and the pre-recorded electronic material. Unlike a traditional performance of an electroacoustic work, which might use a static tape, this piece utilizes a pre-recorded track based on the saxophone score of Jesús Villa-Rojo's *Lamento* (1989) (Villa-Rojo, 1989). This score served as the basis for the Motiv dataset (Carvalho and Sousa, 2025b) used to train the system's underlying motion-based navigation HMM model.

While initial system testing was conducted using the original tape from Villa-Rojo's work, the final performance employs a custom track. To ensure the piece functioned as an original performance rather than a direct recreation, the author remixed the source material (i.e., the original saxophone score and some excerpts from the Motiv dataset) within the Reaper Digital Audio Workstation (DAW) to establish its distinct audio foundation.

12.2.2 Timbre Expansion and Real-Time Modulation

Alongside its navigation features, Aethra employs RAVE models for timbre transfer. This function is implemented in PD, using the `nn_tilde`⁵ external framework developed by IRCAM's ACIDS group. This external allows for the real-time execution of deep learning models directly within the live audio signal path, treating neural synthesis not as a post-process, but as a performative element.

In *Echoes of the Unseen*, we employ three specific pre-trained RAVE models, each providing a distinct topological flavor for the saxophonist to navigate:

- **MusicNet:** Provides a broad palette of orchestral and instrumental timbres, allowing the saxophone to morph into hybrid symphonic textures (Caillon and Esling, 2021).
- **Strengjavera:** Offers specialized string-based and hybrid textures derived from a magnetic resonator piano controlled by artificial life (Armitage, 2023b), facilitating a dialogue between woodwind reediness and autonomous mechanical resonance (Intelligent Instruments Lab, 2023).
- **Aulus-les-Bains Dawn Chorus:** Trained on approximately 230 minutes of environmental recordings by Gregory White, this model serves as the limit of the mutable interface, where instrumental identity merges with naturalistic sound-objects and avian biophony (Intelligent Instruments Lab, 2023).

12.2.2.1 Module Architecture and Latent Weighting

The timbre transfer process begins with inputting the audio signal of the saxophone performer into the RAVE encoder through `nn_tilde` in `encode` mode. The encoder translates the acoustic signal into an n -dimensional latent vector (n is model-specific, e.g., $n = 16$ for MusicNet, while

⁵https://github.com/acids-ircam/nn_tilde, last accessed January 29, 2026.

$n = 8$ for Strengjaverá and Dawn). Then the latent vector is input into the same model’s decoder through `nn_tilde`’s `decode` mode.

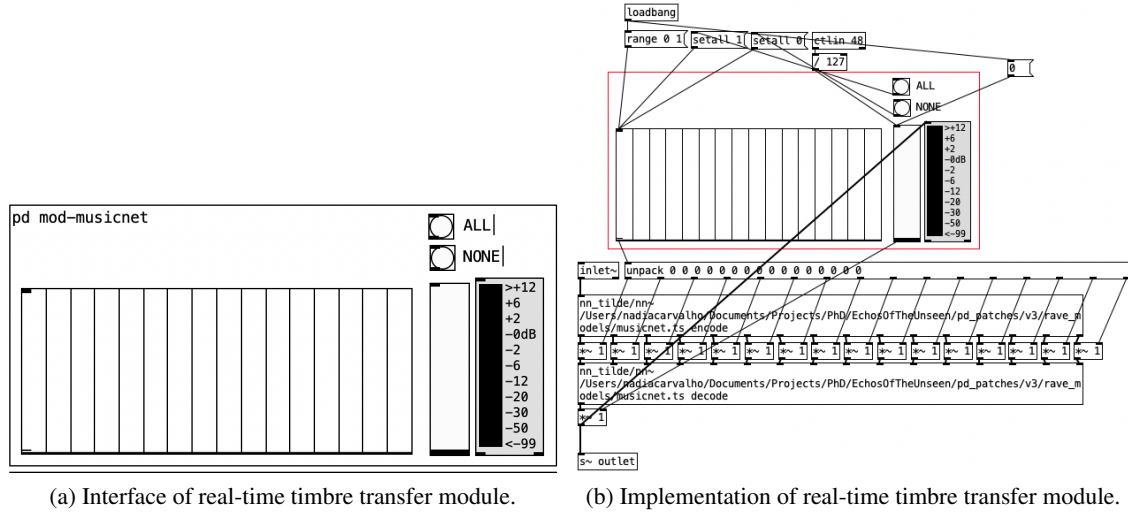


Figure 12.9: Interface (a) and implementation (b) of the real-time timbre transfer module in PD using the `nn_tilde` object with the MusicNet model. The architecture illustrates the encoder and decoder stages of the process, with a 16-channel set of sliders used to modulate latent dimension weights.

A central function of this implementation is the real-time modification of latent dimension weights between the encoding and decoding stages. As shown in Figure 12.9, a bank of n multipliers ($\sim$) intercepts the latent vector, providing variable control over the synthesized timbre through three primary mechanisms. First, by manipulating a multi-slider interface (`multi.vsl`), the performer adjusts the weights of individual latent features. Because the RAVE latent space organizes information hierarchically, this functionality enables the real-time filtering of acoustic parameters, including noisiness, pitch stability, and spectral envelopes. Second, a master intensity parameter controls the overall weight of the latent representations, enabling the performer to transition between the unprocessed acoustic saxophone audio and the synthesized RAVE output. At last, by using the `setall 0/1` logic to mute specific latent dimensions, the performer can alternate between an accurate reconstruction of the source audio and highly altered, non-linear synthesized outputs.

12.2.2.2 Performative Presets and Latent Archetypes

Based on the timbre mapping experiments done in Chapter 9, we established a series of presets that provide the performer with specific reference coordinates within the latent space. The latent weight distributions for these configurations, which vary from accurate audio reconstruction to highly modified spectral outputs, are detailed in Table 12.2.

Table 12.2: Latent Dimension Weight Configurations ($z_1 - z_{16}$) for Performative Presets.

Preset Archetype	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	$z_9 \dots z_{16}$
MusicNet									
Transient Emphasis	1.0	0.7	0.7	0.7	0.3	0.3	0.3	0.3	0.0
Pitch-Shifting	0.8	0.2	0.2	0.2	1.0	0.4	0.4	0.4	0.4
Strengjavera									
Resonant Sustain	0.3	0.9	1.0	0.8	0.5	0.5	0.2	0.2	N/A
Biotic Friction	0.7	0.4	0.2	0.2	1.0	0.9	0.8	0.6	N/A
Dawn									
Biophonic Trigger	0.2	0.2	1.0	1.0	1.0	0.8	0.8	0.8	N/A
Spectral Blur	0.1	0.1	0.4	0.9	0.9	1.0	0.7	0.7	N/A

MusicNet’s “Transient Emphasis” preset targets the first latent dimension to isolate percussive attacks and slap notes. On the other hand, the “Pitch-Shifting” preset concentrates activation on dimension five to generate non-linear pitch-bending effects and microtonal *glissandi*. Strengjavera’s “Resonant Sustain” preset enhances the sound of sustained vibrations by blending soft saxophone tones with synthesized piano string resonance. The “Biotic Friction” introduces rhythmic variations and scraping textures to the audio output. At last, Dawn’s “Biophonic Trigger” preset reduces the weight of harmonic-related dimensions to produce bird-like sounds. Meanwhile, the “Spectral Blur” preset emphasizes deeper, less reactive latent dimensions to convert complex multiphonics into continuous, blurred audio textures.

12.2.3 Improvisational Structure and Formal Score

To guide the free improvisation and provide the performer with a reliable framework, the piece utilizes an “auto-score.” This auto-score outlines specific structural guidelines and assigns defined saxophone techniques, such as unique fingerings and multiphonics, across the circa 10-minute performance, as illustrated in Figure 12.10.

The transition between these different latent archetypes is managed through the same single-knob control, where the *closeness* parameter (ranging from 0 to 1 or using discrete motions) governs a multi-stage linear interpolation across the 16-dimensional latent space.

This parameter organizes the performance into five primary structural zones:

- **0:00-1:08 minutes (The Proximity Zone: $0 \leq \text{closeness} \leq 0.3$ / Parallel motion):** The system targets MusicNet’s “Transient Emphasis” preset to maintain high acoustic fidelity and a direct alignment between the live instrument and the electronic output. From 0:00 to 1:08, the closeness parameter remains near zero while the saxophonist performs stable multiphonics and clean articulations at a pianissimo dynamic. Between 1:08 and 2:48, the performer introduces more complex techniques, such as bisbigliando fingerings and louder dynamics, while the system maintains its low latent deviation.

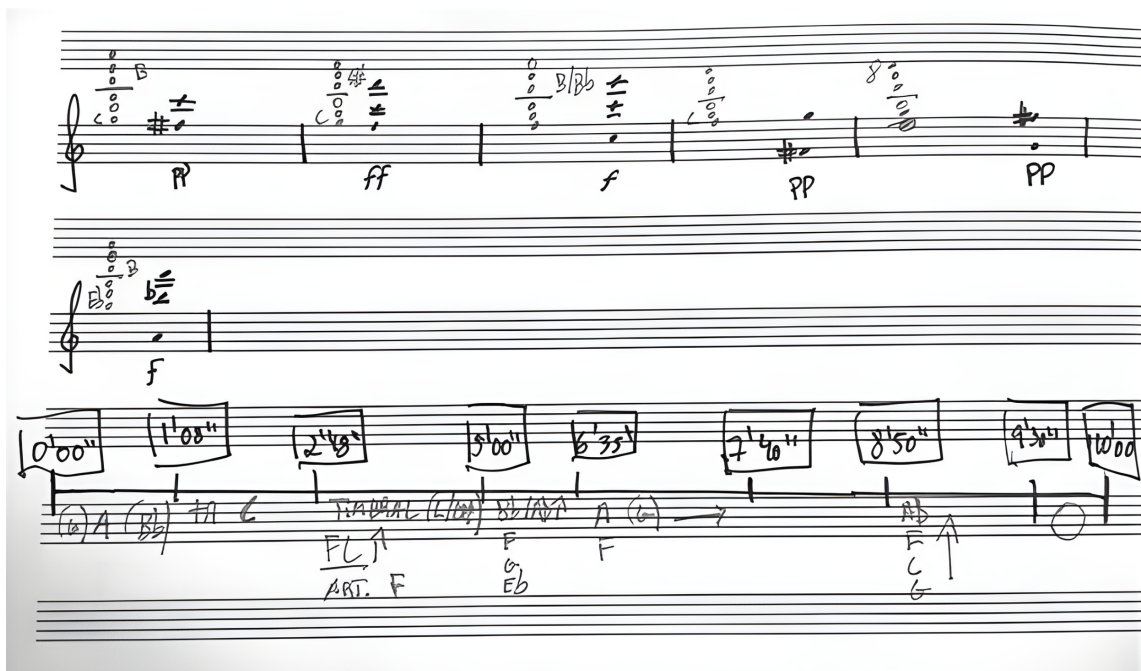


Figure 12.10: Representative excerpt of the *Echoes of the Unseen* auto-score. The notation facilitates a state of contextual navigation by providing a formal skeleton for the performer to explore the high-dimensional latent space, mediating the vulnerability of free improvisation through pre-defined timbre landmarks and extended techniques.

- **2:48–5:00 (The Morphing Zone: $0.4 \leq \text{closeness} \leq 0.7$ / Oblique motion):** The closeness parameter gradually increases, triggering Strengjavera’s “Resonant Sustain” and MusicNet’s “Pitch-Shifting” presets. The use of these presets introduces algorithmic resistance and deviation, generating hybrid piano-string resonances while the saxophonist executes less stable multiphonics at a fortissimo dynamic.
- **5:00–6:35 (Pause and Recalibration):** The system temporarily resets the closeness parameter near zero to re-establish the baseline audio. During this phase, the saxophone part is sparse, utilizing stable multiphonics and alternative fingerings at a reduced dynamic level.
- **6:35–8:50 (Cumulative Generative Phase):** The system switches to the cumulative synthesis mode, processing the saxophone’s high-register altissimo notes and flutter-tonguing into autonomous synthesized textures using the Strengjavera model.
- **8:50–10:00 (The Entropy Zone: $0.8 \leq \text{closeness} \leq 1.0$ / Contrary motion):** The closeness parameter reaches its maximum value at the 9:30 mark. The system activates the Dawn model’s “Biophonic Trigger” and “Spectral Blur” archetypes, muting the primary reconstruction dimensions. This effectively removes the acoustic identity of the saxophone, outputting synthesized environmental sounds to conclude the piece with a final decrescendo.

12.2.4 Visual Synthesis and Latent Mapping in TouchDesigner

The visual component of *Echoes of the Unseen*, developed in TouchDesigner,⁶ displays the high-dimensional latent space of the RAVE models. This setup renders the Aethra system’s algorithmic navigation visible to the audience

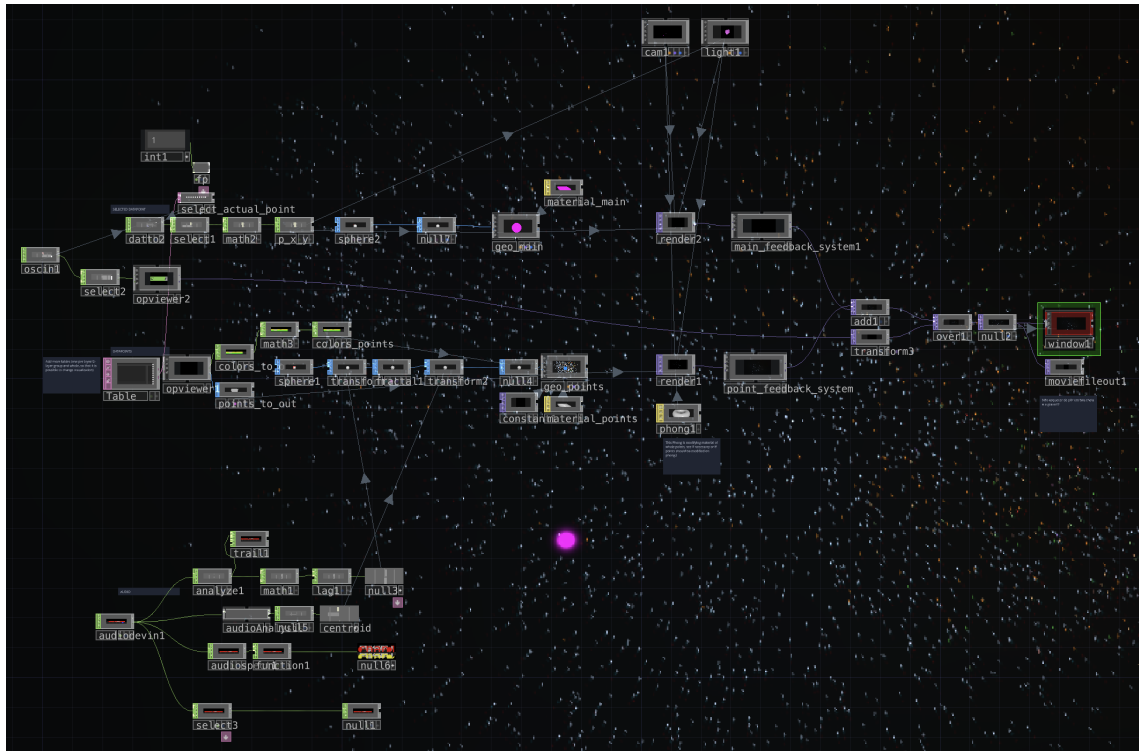


Figure 12.11: *Echoes of the Unseen*’s visual engine in TouchDesigner. The interface displays a point-cloud representation of the latent space, a proximity cursor for real-time navigation, and a morphological feedback slider in the background, as well as the architectural components that comprise the visual system in the foreground.

The TouchDesigner environment (`.toe`) is structured to process data streams from the PD environment using the OSC protocol and audio routing. The architecture consists of three primary functional components. First, `OSC In` and `Select CHOPs` receive the index of the active audio unit and the current *closeness* parameter. These values determine the navigation coordinates displayed within the visual interface. Then, the system uses geometry instancing to display the latent space of the RAVE models. Geometric primitives, similar to spheres and corresponding to each point in the space, are rendered simultaneously using 2D coordinates generated by the UMAP (Uniform Manifold Approximation and Projection) dimensionality reduction algorithm. This data is input via a `Table DAT`, producing a static visual map of the entire timbre dataset. At last, we operate TouchDesigner’s real-time audio processing capabilities (`Analyze` and `Audio-Analysis TouchDesigner` operators) to track the real-time spectral and dynamic characteristics of the live saxophone audio. These extracted features dynamically modulate the width, height, and

⁶<https://derivative.ca>, last accessed January 29, 2026.

luminosity of the rendered geometric primitives to directly reflect the performer's articulation and amplitude.

The interface utilizes specific visual elements to indicate the navigation state between the performer and the system. First, a purple sphere, which we call the proximity cursor, indicates the current frame being played within the latent space. Its movement across the point cloud corresponds to the navigation through the latent units. Second, a green vertical slider displays the current *closeness* value. When the slider is at “C” (Close) indicator, the system is at maximum proximity to the original sequence, accurately reconstructing the source audio. As the value increases toward the “F” (Far) indicator, the system transitions into higher deviation zones, and the visual display becomes increasingly fragmented. Simultaneously, the visual engine applies feedback loops and fragmentation effects (Feedback TOPs) that react in real time to the live audio input. The visual representation of the latent map undergoes rapid blurring and displacement, providing a direct visual correspondence to the spectral modifications applied to the audio signal.

By displaying these internal processes, the TouchDesigner setup renders the neural audio synthesis visible to the audience. This allows observers to visually track how the original saxophone audio is processed, modified, and reconstructed as the performer interacts with the system's defined parameters.

12.2.5 Dissemination and Performance History

The application of the Aethra system as a co-creative performative tool was evaluated through a series of international performances and conference presentations of *Echoes of the Unseen* (see Table 12.3). These events facilitated an iterative development process, enabling refinement of the system's design to meet the operational demands of live improvisation. To supplement this written analysis, complete audiovisual documentation of all *Echoes of the Unseen* performances can be accessed at the project's digital repository, <https://nadiacarvalho.github.io/MutableMusicalInterface/#/software/aethra>.

During initial evaluations at the IX Encontro Internacional da Casa das Ciências and the Conservatorio Superior de Música de Castilla-La Mancha, a preliminary version of the system utilizing distance-based navigation was tested. Observations indicated that without a formal notational framework, the performer encountered difficulties anticipating the system's output, as navigation relied entirely on intuitive interaction with the high-dimensional latent space. Consequently, these early trials identified two primary requirements for optimizing the co-creative environment: a more precise mechanism for interpreting performative gestures and a formal structural framework to guide the performer's intentions.

Subsequent performances at the Audio Mostly & ICAD Joint Conference (AM.ICAD) and FISP events utilized an updated system featuring motion awareness. This version already implemented HMMs as a navigation strategy, enabling the system to actively anticipate performative gestures rather than strictly reacting to inputs, and introduced the initial auto-score. By the performance at Auditorio Carlos do Carmo, the system reached its current configuration by integrating the single-knob control paradigm and the TouchDesigner visual interface to facilitate contextual

Table 12.3: Performance and Dissemination History of *Echoes of the Unseen*.

Date	Event / Venue	Location
8 July 2024	IX Encontro Internacional da Casa das Ciências	Aveiro, Portugal
1 April 2025	“La interpretación como medio creativo en la era post-digital,” Conservatorio Superior de Música de Castilla-La Mancha	Albacete, Spain
1 July 2025	Audio Mostly & ICAD Joint Conference (AM.ICAD), Casa das Artes de Coimbra	Coimbra, Portugal
7 July 2025	FISP - Festival Internacional de Saxofones de Palmela	Palmela, Portugal
29 Nov 2025	ENIM - Encontro Nacional de Investigação em Música	Évora, Portugal
11 Feb 2026	“Ressonâncias do Presente: música para saxofone e multimédia,” Saxophone Recital, Auditorio Carlos do Carmo	Lagoa, Portugal
Late July 2026	Professional Studio Recording and Documentation	Aveiro, Portugal

navigation of the latent space. Iterative modifications across these venues established Aethra as a stable, mutable interface. As of early 2026, the system’s development phase is complete, and a professional recording of *Echoes of the Unseen* is scheduled for late July 2026 in Aveiro, Portugal.

12.2.6 Creative Affordances of Aethra in Performative Context

The evaluation of Aethra through the *Echoes of the Unseen* case study details its impact on contemporary music-making by transforming the electronic medium from a static backing track into a responsive, co-creative partner. This section evaluates the performer’s interaction with the Aethra system, outlining the observed benefits, limitations, and the impact of navigation strategies on performance execution. The analysis subsequently addresses how the system maintains musical identity while operating within algorithmic material resistance and defined aesthetic constraints. Finally, the evaluation reviews the application of timbre transfer presets and their structural effect on the musical output. Through this framework, the section assesses how Aethra transitions the performer’s role from interpreting a fixed score to functioning as a post-performer within a responsive digital environment, highlighting the system’s broader impact on contemporary music making and its capacity to transform static works into mutable, co-creative interfaces.

12.2.6.1 Navigation Strategies and Synthesis Logic

The transition from strictly distance-based mapping to motion-aware navigation constitutes a significant advancement in the system’s operational framework. During early iterations, Aethra relied on proximity-based methods that yielded highly predictable interactions, causing the performer to experience the latent space as a static, predetermined topography rather than a responsive entity. This limitation was particularly evident at lower deviation thresholds, closer to the original *closeness* values, where the interaction could occasionally feel rigid or frozen.

The integration of the motion-aware HMM model resolved this limitation by enabling the system to evaluate the trajectory of gestural inputs, transforming the interface from a passive retrieval tool into a proactive musical agent. By interpreting the underlying gestural data of the fixed electronic tape, the motion-aware strategy allows the system to generate coherent linear responses. For instance, when utilizing parallel motion, the generated output maintains a much higher degree of structural cohesion with the original material while successfully integrating the anticipated variation. Similarly, during contrary motion, the system produces a linear, intelligible output that facilitates a reciprocal musical dialogue, an outcome that was frequently disjointed under the purely distance-based paradigm.

This navigational shift fundamentally redefines the composition from a static artifact into a continuous manifold explored through intentional thematic development rather than discrete database querying. By replacing random drift with intentional trajectory mapping, the system establishes a framework where authorship is distributed between the human performer's physical intent and the machine's contextually bounded possibilities.

Furthermore, this development directly influences audience perception. The visual proximity cursor, which previously mirrored direct user control, begins to exhibit stylistic inertia within the visual interface. This visual feedback explicitly communicates the machine's independent agency, shifting the observer's understanding of the performance from a unidirectional control mechanism to a distributed, co-authored dialogue.

Finally, the system's dual-mode synthesis logic further augments the performer's compositional agency, functioning interdependently with the performance's overall formal structure to dictate whether the focus remains on acoustic reconstruction or structural evolution.

The linear fidelity mode sustains a direct correlation with the original acoustic baseline, effectively preserving the instrument's core identity. In this mode, the playback system generates a parallel continuous output that operates alongside the live saxophone, remaining active even when the *closeness* parameter is set to higher deviation thresholds or when contrary motion is utilized. This mode is specifically prioritized during the Proximity Zone (0:00–2:48) to firmly establish a clear connection with the instrument's acoustic identity for the audience.

Conversely, as the performance transitions into the Entropy Zone, the compositional focus shifts from accurate reconstruction to necessary structural variation. Because the linear fidelity mode cannot provide the required level of deviation, the system switches to the cumulative transformation mode. This engine aggregates disparate latent segments to generate highly non-linear, high-entropy variations. This algorithmic shift directly impacts the human performer, encouraging the saxophonist to adopt a more unconventional and unstructured improvisational approach, thereby actively enhancing the complete deconstruction of the original musical work.

12.2.6.2 The Performer's Vulnerability and the Auto-Score

For classically trained musicians rooted in the deterministic Western score-based tradition, the transition to free improvisation within an AI-mediated latent space introduces significant performative vulnerability. During initial evaluations of the system (performance at the IX Encontro

Internacional da Casa das Ciências), the absence of a formal notational framework forced the performer to rely solely on intuition to navigate the complex sonic environment. Observations and performer feedback from these early trials indicated a profound sense of discomfort and risk. Without a predetermined timeline or structured guide, the musician struggled to anticipate the system's generative responses and felt exposed to the unpredictability of the machine. This performative vulnerability is consistent with recent findings (Benford et al., 2024), who emphasize the complex negotiation of trust and autonomy required when musicians engage with unpredictable AI partners.

To mitigate this uncertainty and restore the performer's interpretative agency, an auto-score was developed to provide necessary structural support. The development of this auto-score was centered specifically around the elements the performer felt were most necessary to include. Rather than relying on traditional melodic notation, the performer needed concrete, tangible acoustic landmarks to confidently interact with the AI. Consequently, it became necessary to include a flexible set of specific sound techniques that the saxophonist could rely on as an anchor (e.g., specific multiphonics, unique fingerings, and defined dynamic envelopes).

By explicitly including these specific instrumental playing techniques, the auto-score provided the exact structural scaffolding the performer required to navigate the 10-minute performance without feeling lost. The implementation of this framework successfully transformed the performer's initial discomfort into a productive, creative friction. With the auto-score in place, the musician no longer felt the vulnerability of completely free improvisation. Instead, they could utilize the provided landmarks to comfortably align their musical gestures with the shared goals of the system, establishing a reciprocal dynamic where the human and machine actively inspire one another without collapsing into disjointed noise.

This process highlighted that the work exists only in the interaction. In this light, the performance is a snapshot of a broader development where the performer helps shape the machine's intentionality, and the machine challenges the performer to adopt textures and registers that would be instrumentally unreachable without this digital interlocutor.

12.2.6.3 From Interaction to Co-Creativity and Contextual Navigation

A critical finding in this case study is the achievement of a player paradigm, where the machine acts as a reactive interlocutor rather than a passive effect processor. The creative affordance of Aethra lies in its ability to extend the performer's agency beyond human physical limits. As noted by the performer:

The interest is not in hearing the machine replicate the gesture... What becomes truly compelling is when the machine uses a *bisbigliando* in unexpected ways: morphing it into a multiphonic or a trill shaped like a *glissando*.

However, the path to achieving this seamless contextual navigation required navigating significant technical hurdles during the system's iterative development. During the AM.ICAD performance, the motion-aware system operated correctly and stably, primarily because the setup was restricted to the PD audio environment without the simultaneous rendering of visuals.

In contrast, the intermediate version of the Aethra system, deployed subsequently at the FISP and ENIM festivals, exposed profound computational bottlenecks. For these performances, the multimodal TouchDesigner visual engine was connected to run concurrently with the PD audio synthesis. Because the system was operating on a previous, less optimized version of the py4pd library, and the real-time neural timbre-transfer modules had not yet been fully optimized for such high-intensity parallel processing, the sheer computational load overwhelmed the architecture. This resulted in noticeable audio clipping, latent lag, and general system instability during the live improvisations.

Yet, within the theoretical framework of the Aethra system, both these early technical struggles and the inherent unpredictability of the high-dimensional RAVE space are treated as necessary material resistance. This resistance manifests as a non-linear response to input, causing the sound to drift or morph in unexpected ways. Rather than viewing these unpredictable, non-linear sonic responses as technical flaws, the framework actively reframes them as creative friction. This resistance forces a departure from the perfect transparency usually sought in digital instruments, demanding a specific "post-digital expertise." As the saxophonist noted, performing with the machine requires a fundamental shift in auditory perception, in which musicians must "close their eyes" and deeply listen to the system's generative output, learning to treat its unique acoustic quirks as new musical ideas rather than mistakes. By doing so, they build a reciprocal relationship with the AI, embracing its unique material resistance just as they would with any traditional acoustic instrument.

To successfully manage this complex material resistance without overwhelming the musician, Aethra relies heavily on its one-knob control paradigm. By compressing all high-dimensional latent information into a single performative axis, the system achieves high control intimacy. This intentional simplification allows the performer to focus entirely on the aesthetic relationship between their physical gesture and the resulting timbre shift, rather than managing software technicalities.

Ultimately, the performer navigates the machine's material resistance by sensing the boundaries of the training data. This ongoing feedback loop establishes an entangled performative network, situating the true creative affordance of Aethra in the continuous, interdisciplinary dialogue between informatic logic and embodied human intuition. At the same time, identifying and resolving these specific py4pd and timbre-transfer bottlenecks was essential. It allowed the later, mature version of Aethra to move past technical survival and achieve true, fluid co-creativity, solidifying the musical work not as a static artifact, but as a continuous, emergent evolution shaped by human intent and digital mediation.

12.2.6.4 From Manual Control to Contextual Navigation: Neural Archetypes as Performative Landmarks

The use of specific latent weight configurations, or latent archetypes, emerged as a crucial solution through the project's iterative development cycle. In early performances of *Echoes of the Unseen*, the sheer complexity of manipulating the 16-dimensional RAVE space necessitated a second technician to manually control the electronic parameters and operate the system's interface. This reliance on an external operator divided the performative agency, inherently distancing the saxophonist from a direct, embodied connection with the digital system's generative responses.

To resolve this limitation and fully liberate the performer from technical assistance, the final mature version of the system introduced curated performative presets (see Section 12.2.2.2). These presets function as curated snapshots of the machine's learned internal logic, designed to expose specific affordances of the mutable interface without requiring manual multidimensional mixing. By surgically weighting dimensions (e.g., maximizing z_1 for energy tracking or z_5 for pitch-shifting), the system translates complex neural mappings into stable, predictable aesthetic landmarks, enabling the performer to negotiate with the model's material resistance in a highly intentional way.

The implementation of these archetypes successfully consolidated control intimacy directly into the hands of the musician. With the high-dimensional filtering automated by the presets, the performer no longer needed a technician and could focus entirely on operating the single-axis control knob. Furthermore, the distribution of these archetypes across the one-knob control paradigm facilitates a specific form of control intimacy. The division into Proximity, Morphing, and Entropy zones allows the ensemble to manage the complexity of a 16-dimensional manifold through a single parameter. This simplification is critical for the post-performer, as it drastically reduces cognitive load, allowing the musician to focus on contextual navigation rather than the technicalities of the software. The presets thus act as functional anchors that help the performer navigate a complex and highly unpredictable digital model.

Ultimately, the successful deployment of these presets demonstrates that the musical work functions as a dynamic ecosystem. The ability of the performer to intervene in the *closeness* parameter in real-time reinforces the intra-active nature of the work, where the final sonic outcome is generated by multiple agents interacting within the same computational environment. This iterative advancement transformed Aethra from a technician-operated software patch into a true post-digital instrument, empowering the musician to seamlessly navigate the work's topological universe in real time.

12.2.6.5 The Visual Interface as a Cognitive Bridge

In live AI-mediated music, a persistent perceptual gap exists where audiences often struggle to discern whether a sonic event is the direct result of human action, autonomous machine generation, or fixed media playback (Naseck et al., 2025). To address this opacity in *Echoes of the*

Unseen, the visual layer developed in TouchDesigner functioned as a critical cognitive bridge, actively mediating the relationship between the system's hidden algorithmic processes and audience perception.

By projecting the high-dimensional RAVE latent space into a legible 2D topological map, the visual interface explicitly addressed this perceptual gap. It translated complex neural computations into a comprehensible spatial environment, allowing the audience to actively track the machine's real-time decision-making processes and observe its material resistance.

Central to this visual narrative is the proximity cursor (represented by a distinct purple sphere), which maps the system's active navigation across different acoustic clusters. Rather than merely reacting to the saxophone's pitch, the cursor's trajectory visualizes the machine's algorithmic search through the data. Smooth, continuous movement indicates a state of high-fidelity reconstruction, whereas erratic, fragmented jumps visually expose the system operating in a high-entropy, generative state. This erratic movement effectively demonstrates to the audience the machine's active attempt to negotiate with the performer's musical intent.

Furthermore, the interface establishes interactional transparency through dedicated visual feedback mechanisms. As the performer adjusts the *closeness* parameter, a morphological slider provides real-time, quantitative feedback, visually indicating the exact degree of deviation from the original tape. Simultaneously, the visual engine employs feedback loops that mirror the acoustic deconstruction of the saxophone's timbre. As the sound shifts from stable tonal structures to scattered, chaotic textures, the visual point cloud fragments correspondingly, explicitly linking the performer's gestural input to the digitally-enabled sonic deconstruction.

The necessity of this multimodal transparency for true co-creativity was particularly evident during the performance's semi-silent phase (between 5:00 and 6:35). As the acoustic dialogue became sparse, the visual interface snapped into sharp focus, re-anchoring the audience to the core topological elements of the piece.

Ultimately, the visual interface functions as an essential epistemic guide, elucidating the tension between human intentionality and algorithmic agency. It transforms the opaque black box of deep generative models into a shared, legible space, inviting the audience to witness the dynamic, mutable entanglement of the performer and the machine.

12.3 Summary: Aethra, A Mutable Interface for Mixed Music Navigation

The exploration of *Echoes of the Unseen* through the Aethra system provides a practical validation of using motion-aware latent space navigation as a framework for live improvisation. By integrating RAVE models within a PD (via py4pd) and TouchDesigner ecosystem, the architecture successfully addresses the black box problem inherent in deep generative models. The technical capacity to manipulate 16-dimensional latent weights in real time, coupled with a transparent visual latent map, transforms abstract computational data into a navigable creative territory. This

paradigm shift effectively redefines the role of the contemporary musician, evolving them from an interpreter of fixed symbols into a post-performer who actively navigates topological variety.

While the system's deployment across multiple venues confirmed its artistic value and robustness, it also highlighted critical areas for future development to strengthen the mutable interface. To further deepen the state of contextual navigation, future iterations must prioritize enhanced machine listening and mimicked agency. By implementing predictive computational models that are highly sensitive to micro-temporal changes in dynamics and articulation, the system can better anticipate the performer's trajectory. Refining this control intimacy will ultimately reduce latent lag and decrease the performer's reliance on physical, knob-based interaction, fostering a more direct, embodied connection to the digital sound-body.

Additionally, the technical complexity of the py4pd and RAVE integration remains a significant barrier to wider accessibility. Future developments aim to streamline and simplify the pipeline for training and curating customized datasets. This democratization would empower performers without specialized programming expertise to treat the latent space as a personalized plastic material, allowing them to independently define the material resistance and timbre boundaries of their own performative ecosystems.

Finally, building upon the successful implementation of the TouchDesigner visual engine, subsequent research should investigate more immersive visualization techniques, such as Augmented Reality (AR) and more granular visual feedback. Integrating these technologies would allow both the performer and the audience to perceive the high-dimensional connective tissue of the sound in far more direct, spatialized ways.

Ultimately, Aethra represents a significant step toward making technology an essential, co-creative musical partner rather than a passive execution tool. It bypasses the limitations of fixed tape music by transforming the static electronic backing track into a responsive collaborator, doing so without compromising or overwriting the composer's original work.

By shifting from fixed audio playback to dynamic latent space navigation, driven by an AI-informed concatenative synthesis (CBCS) framework and a granular playback engine, the system algorithmically adapts to the musician's live gestures in real-time. Crucially, rather than synthesizing new waveforms from scratch that might alter the piece's identity, the system dynamically retrieves and concatenates the actual raw audio recordings associated with the selected latent indices. This design ensures that the absolute acoustic fidelity, morphological richness, and authentic expressive nuances of the composer's source material are rigorously maintained.

This enables the machine to generate continuous, evolving sounds without relying on a pre-determined timeline, effectively restoring expressive and stylistic agency to the performer while honoring the structural integrity of the original composition. By establishing this strong foundation of motion-aware interaction, this research paves the way for the next generation of collaborative audiovisual experiences.

Synthesis to Part III:

Timbre, Gesture, and Motion in Latent Spaces

Part III has extended the investigation from structural aspects of music to its expressive dimensions, timbre, gesture, and motion. Through latent space models, these qualities have been analyzed, represented, and integrated into interactive systems, enabling new forms of co-creation in both compositional and performative contexts. The studies on the timbre landscape, the Motiv dataset of motion, and motion-enhanced sampling collectively illustrate how expressivity itself can become navigable and generative.

The foundation of this research, in Chapter 9, lies in the RAVE model’s ability to map the tenor saxophone’s complex sonic palette into a 16-dimensional latent space. Initial studies have demonstrated that this space effectively clusters dynamics, pitch registers, and extended techniques, with specific dimensions showing heightened sensitivity to certain acoustic features. For instance, dimension 10 aligns with pitch, and dimension 15 is related to extended techniques.

Building upon this landscape, Chapter 10 explored navigation strategies that move beyond stochastic sampling toward intentional manipulation. By implementing sampling methods such as *k*-nearest neighbors (*k*-NN) for equal-density binning and Distance-Based (BD) for spatial division, we introduced a control parameter *C* that allows performers to calibrate their proximity to a target sonority. While lower values of *C* maintain the integrity of the source material, higher values allow for controlled deviation, although experiments show that high values often necessitate trajectory alignment to maintain musical coherence.

To bridge the gap between abstract latent dimensions and musical intent, Chapter 11 adapted a taxonomy of parallel, oblique, and contrary motions to categorize trajectories within the manifold. Analysis revealed that parallel motions tend to preserve the strongest structural similarity by following the original path through transposition, whereas contrary motions represent radical divergence characterized by significant spatial dilation and lower structural correlation. These geometric patterns provide an interpretable structure for the latent space, transforming it from a static repository into a navigable terrain of gestural transformations.

This synthesis culminates in Aethra, a co-creative system developed in Pure Data that empowers musicians to interact with fixed electronic tapes as dynamic partners. By integrating sound recognition, distance-based sampling, and motion-aware navigation into an intuitive one-knob interface, Aethra abstracts the underlying high-dimensional complexity for the performer, allowing

for seamless control. Chapter 12 details the design and development of Aethra and discusses a performance case study *Echoes of the Unseen*. In this case study, the system enables a symbiotic relationship where the acoustic and digital realms evolve together in real-time co-improvisation. This reinforces the concept of the mutable work as a flexible system that thrives on interaction, preparing the ground for the final vision of co-creative electroacoustic practice.

In summary, the transition from analyzing timbre signatures to implementing Aethra marks a shift from viewing AI as a reproductive tool to a medium for active performative appropriation. The expressive latent space functions as a topological map where musical intent, encoded as directional motion, is navigated in real-time. By formalizing the relationships between original phrases and their parallel, oblique, or contrary variations, this research establishes a motion-aware sampling framework that maintains stylistic coherence despite deviations from source material.

The concluding insights of Part III suggest that co-creative practice thrives on balancing structural fidelity with expressive deviation. While distance-based navigation can lead to dissonance at high divergence values, the integration of trajectory alignment and motion identification ensures the generative response remains musically grounded and contextually aware. By distilling high-dimensional navigation into intuitive gestural controls, like the one-knob interface used in *Echoes of the Unseen*, these developments bridge the gap between neural representations and the lived experience of performance, framing the mutable work as a flexible system driven by human-machine interaction. These findings set the stage for Part IV, where tonal and expressive frameworks merge into a comprehensive vision for AI-mediated electroacoustic music.

Part IV

Towards a Co-Creative Practice: Synthesis and Conclusions

Introduction to Part **IV**:

Towards a Co-Creative Practice: Synthesis and Conclusions

While Parts **II** and **III** provided the artistic, technical, and experimental grounding for navigating tonal and timbre manifolds, Part **IV** serves as the final synthesis. It evaluates the shift from the musical work as a fixed artifact to a mutable interface. This part addresses the research questions by integrating the findings from the BroadcastJSB and Aethra systems, discussing the ontological implications of contextual navigation, and providing a final reflection on the role of the composer-performer in an age of intelligent mediation.

The part is organized into two main chapters. The first, Chapter **13** presents a synthesis organized by the core research questions posed in Chapter **3** and attempts to draw important insights from the presented investigations to articulate the role of machine learning systems as collaborative partners in composition and performance, focusing on the transformation of the musical work to a mutable interface. At last, Chapter **14** concludes our work by highlighting the contributions of this research to multiple fields, acknowledging its limitations, and outlining directions for future work.

Chapter 13

Synthesis and Discussion of Findings

This chapter looks at the research results in relation to the main questions we set out to answer in Chapter 4. It explores what the findings of Parts II and III mean for modern electroacoustic practice. The discussion is centered on two main ideas: 1) the ontological construction of the mutable work through structured representation and navigation, and 2) the performative negotiation of agency within co-creative systems. Table 13.1 summarizes the findings and facilitates the reading of the chapter.

13.1 Ontology and Navigation of the Mutable Work (Addressing RQ1)

A key idea of the mutable interface is the shift of focus to topological manifolds as the locus of musical representation. R1 questions:

Can high-dimensional latent spaces be structured to reflect perceptual, mathematical, and music-theoretical concepts, and what navigational paradigms allow these manifolds to function as intelligible, mutable musical works?

We consider this through two such perspectives: the structural logic of the manifold itself (SQ1.1) and new navigation approaches (SQ1.2) that abstract the database lookup into geodesic traversal in high-dimensional space.

13.1.1 Structural Logic (SQ1.1)

To investigate the structural logic of the manifold itself, SQ1.1 questions:

Are unsupervised latent spaces empirically aligned with formal music theory and perceptual sound characteristics to ensure that the mutable work maintains stylistic tonal coherence and timbre intelligibility?

The research found that we can connect unsupervised learning models to formal music theory. First, the research results (Chapter 5) show that the type of encoding used in training VAEs acts

Table 13.1: Synthesis of the empirical and performative findings of this research, organized by the two primary research pillars and the dual domains of Tonal and Timbre logic.

Sub-Question	Domain	Core Synthesis Finding
SQ1.1: Representation & Logic	Tonal	Unsupervised latent spaces (MusicVAE) discover formal tonal logic (e.g., the circle of fifths). This is demonstrated from the alignment with spectral pitch and pitch class distributions (DFT-P and DFT-PC), achieving correlations of $\rho \approx 0.65$ and $\rho \approx 0.61$, respectively. They prioritize harmonic function (semantic) over literal pitch retention, existing in a state of hierarchical tension where the manifold reflects parsimonious voice-leading but clarifies functional clusters through the statistical grammar of the corpus (e.g., $I - V$ proximity over $I - ii^7$).
	Timbre	Latent spaces (RAVE) autonomously cluster extended techniques, creating a perceptually intelligible taxonomy based on spectral density and morphological similarity (SaxTimbre).
SQ1.2: Navigational Paradigm	Tonal	Shifting from discrete to topological querying (k -NN with angular displacement) preserves tonal momentum, allowing identity to persist through structural fluidity.
	Timbre	Mapping high-dimensional trajectories to counterpoint taxonomies (parallel, oblique, contrary) transforms raw navigation into intentional, musically intelligible thematic development.
SQ2.1: Epistemic Tools & Agency	Tonal	Real-time exploration of harmonic manifolds via BroadcastJSB redistributes authorship, shifting the performer's role from executing a fixed score to navigating an epistemic territory.
	Timbre	The high-dimensional entanglement of neural manifolds in Aethra provides the necessary material resistance to transform the machine from a passive tool into an active performative agent.
SQ2.2: Contextual Navigation	Tonal	Constrained tonal navigation through latent spaces enables meaningful contextual navigation, where the machine's contextually bounded possibilities interact dynamically with human stylistic intent.
	Timbre	Compressing timbre complexity into motion-aware, one-knob XAI interfaces fosters contextual navigation, allowing performers to internalize the machine's contextually bounded possibilities as an aesthetic property.

like a filter for how the AI understands music. Text-based encodings, like ABC notation, are great for accurately recreating music, with an accuracy of 82.6%. However, they do not effectively organize musical information in a way that makes sense from a music theory perspective. On the other hand, using the DFT from pitch distributions (DFT-P) as an encoding performed much better at capturing cognitive tonal hierarchies. Specifically, in 16% of the J.S. Bach chorales tested, the model achieved a near-perfect circular correlation (Kendall's $\tau \geq 0.9$) between the ordinal sequence of keys in the latent space and the theoretical circle of fifths. This highly organized subset consisted entirely of major-key and longer compositions (exceeding 114 slices), implying that when given sufficient tonal evidence, DFT-P accurately reflects Western perceptual musical principles (e.g., circle of fifths).

Furthermore, a key discovery was that MusicVAE mostly captures tonal structure through the angle of its vectors rather than their size (see Section 6). This aligns well with established music theory concepts. Particularly in how pitch distributions behave in Fourier Phase Spaces, showing a moderate correlation ($\rho \approx 0.65$). The results also point to the latent space existing in a state of hierarchical tension. While it remains sensitive to the proximity of parsimonious voice-leading, the bottleneck of the VAE forces a prioritization of semantic invariants, most notably harmonic function. As illustrated in the case study of Chorale *BWV 307* (see Section 6.3.1), the latent space effectively functions as a syntactic filter, generalizing tonal functions more clearly than raw DFT spaces by incorporating the statistical likelihood of harmonic progressions (such as the high prevalence of *I – V* transitions). Consequently, the latent space improves the representation of tonal music with a better understanding of the structure of the training data.

This alignment implies that deep learning is not as opaque as it seems. Instead, when trained on highly reinforced structural frameworks like common practice Western music, it can autonomously uncover foundational patterns (e.g., the circle of fifths, functional harmony, and parsimonious voice-leading). Consequently, we can use these latent spaces as representation spaces for this specific tonal language, allowing them to be navigated to generate music from a stylistically consistent corpus. This challenges the traditional notion of *werktreue*, because now the identity of music is maintained through its broader structure. We can maintain topological integrity and move on from strictly using fixed notation.

This unsupervised alignment extends beyond pitch hierarchies into the intrinsic acoustic properties described in Chapter 9. When studying the timbre qualities of the tenor saxophone using RAVE pre-trained models, the research shows that most models do not just collect random audio features. Instead, they autonomously organize these features into semantically meaningful clusters. Specifically, similar sounds (such as specific extended techniques) are grouped together based on their spectral density and morphological similarity, forming a perceptually intelligible taxonomy.

Just as the MusicVAE discovered the circle of fifths without explicit instruction, the RAVE model autonomously aligns perceptual characteristics of sound. Extended techniques, such as multiphonics, flatterzunge, and air sounds, are effectively grouped in RAVE's latent space into distinct regions based on their spectral density. This creates a structural logic for timbre that is

analogous to the functional clusters of Bach’s chorales. In both cases, the neural network acts as an epistemic filter, compressing high-dimensional complexity into an intelligible territory. This proves that structural coherence in a mutable interface is not restricted to Western notation but is a property of the manifold’s ability to map the underlying material logic of any given corpus, whether it be harmonic or timbral in nature.

13.1.2 Navigational Paradigm (SQ1.2)

Addressing new navigation approaches, SQ1.2 questions:

In what ways does the shift from discrete database querying to intent-aware topological navigation redefine the identity of a composition as a continuous manifold rather than a collection of fixed points?

The research postulates that exploring music identity within high-dimensional generative models must evolve from simple linear interpolation into an intent-aware navigational paradigm. Navigating via straight-line mathematical averages assumes a Euclidean geometry that fails to capture the non-linear distribution of musical syntax. This purely mathematical approach frequently forces the system into latent dead zones, generating musically incoherent artifacts or hallucinated chords that lack functional identity. To bypass this continuous noise and synthesize stylistically valid responses, Chapter 7 introduces intent-aware querying strategies for sampling in the tonal latent space of MusicVAE, specifically k -nearest neighbors (k -NN) and k -NN with angular displacement (k -NN+AD). By analyzing the performer’s directional intent, these novel strategies allow the system to sample discrete, perceptually coherent musical events directly from the established corpus.

The empirical findings directly demonstrate that these intent-aware strategies significantly outperform discrete linear interpolation in preserving the stylistic and structural grammar of the training corpus, like the harmonic changes in the JSB corpus. Specifically, standard k -NN successfully maintains local harmonic validity, excelling at preserving the functional integrity of the bass line and parsimonious voice-leading. However, the true novelty of this research lies in the application of k -NN+AD. By calculating the angular trajectory between chords, k -NN+AD effectively captures tonal momentum. This ensures that the upper voices follow parsimonious voice-leading rules and that the broader cadential direction of the phrase is preserved. These results advance the state-of-the-art in developing Western common-tone practice-aligned co-creative systems by proving that the geometry of a latent space (specifically, vector angles) directly encodes musical meaning. Consequently, this shifts the contemporary paradigm from treating AI as an opaque black box generator to utilizing it as a transparent, navigable interface where algorithmic precision and human compositional intent can seamlessly interact.

On the other hand, in the timbre domain, navigation faces the high-dimensional sensitivity of the RAVE manifold. The timbre realm holds the technical discovery that efficient navigation of high-dimensional manifolds requires a shift from geometric nearness to intention-aware gestural logic. Taking advantage of methods such as k -NN for equal-density binning and using a closeness

control parameter (C), we can finely tune the distance of musical samples with respect to different timbre attributes. While not always required, extending k -NN with trajectory alignment ensures stability in the sampled music.

This navigation model is enhanced by a motion-aware taxonomy, comprised of parallel, oblique, and contrary motions (see Chapter 11). This taxonomy was materialized in the Motiv dataset. This approach turns the complex latent space into a more meaningful space for musical exploration. As a result, the research shows that when the structure aligns with these varied movement styles, the interface enables a state of contextual navigation. It allows performers to see movement paths not just as numbers, but as understandable gestures that hold musical significance. For example, the results showed that contrary motion trajectories within the latent space were perceived by performers as exhibiting the highest degree of mimicked agency, as they provided a clear counterpoint to the performer's input.

Comparing the findings of Chapter 7 (tonal) and Chapters 10 and 11 (timbre), it becomes evident that while these strategies provide structural stability, the Motiv investigation introduces a superior navigational layer of musical intent. Crucially, this expands the traditional navigation paradigms of corpus-based concatenative synthesis (CBCS). While conventional CBCS typically relies on static, point-by-point proximity matching within descriptor spaces, often leading to a disconnect between the statistically retrieved audio units and the continuous morphology of the performer's musical gesture-motion-aware sampling treats the corpus as a dynamic topography of gestures. Unlike techniques like k -NN or k -NN+AD, which only respond to the current situation, motion-aware navigation looks deeper into the nature of how a performer interprets and develops thematic material over time. For instance, a contrary motion can trigger an opposing action, allowing the interface to react dynamically rather than being a static map. In the end, this comparison shows that a good navigational strategy needs to balance two important things: how well it keeps the structure of the music and how well it supports the intended performance. While focusing on tone and timbre distances helps make the mapping more straightforward, a navigation method that considers motion allows for better understanding. This ensures that shifts between different musical states are not just close in a mathematical sense but also make sense musically and connect with the performer's intentions, effectively transforming CBCS from a static mosaic engine into a responsive, co-creative instrument capable of genuine musical dialogue.

13.2 On Practical Applications and Contextual Navigation (Addressing RQ2)

To address the performative dynamics and the redistribution of agency within the interface, RQ2 questions:

In what ways does the evolution of the epistemic tool (Magnusson, 2009) redistribute authorship, and how can contextual navigation be fostered through the machine's material resistance?

To answer this, we must examine how the epistemic tool fundamentally redistributes musical agency and facilitates a model of distributed authorship. We will discuss how this interface dismantles traditional, human-centric notions of creative ownership, transcending the conventional tool-user paradigm to establish an entangled network where the human performer and the machine's mimicked agency engage in a continuous dialogue. Drawing from the empirical findings of the Aethra (and its performative study case *Echoes of the Unseen*) and BroadcastJSB case studies, we will analyze the mechanisms through which neural latent spaces impose a productive material resistance that necessitates contextual navigation.

13.2.1 Material Resistance and Epistemic Action (SQ2.1)

Focusing on the epistemic role of the interface, SQ2.1 questions:

How does the implementation of real-time exploration interfaces transform the user's interaction with musical logic from a task of execution to an epistemic process of discovery and navigation?

The high dimensionality and entanglement of neural latent spaces introduce material resistance into the human-machine interaction. However, it is crucial to critically reflect on the ontological status of this AI-based machine within this humanistic domain. The system is not an autonomous, sentient co-creative partner. Rather, it fundamentally remains an epistemic tool. The AI does not possess actual sentience. Instead, it generates epistemic friction through the non-linear density of its latent space, offering structural responses based on the statistical distributions, internal logic, and inherent biases of its training corpus.

While this encoded contextual knowledge facilitates a dynamic dialogue, it also introduces significant artistic limitations and performative challenges that must be questioned, particularly regarding its adoption by other performers. Because the system's material resistance is non-linear and its responses occasionally opaque, performers deeply rooted in the traditional, score-based paradigm often experience a profound state of vulnerability and a perceived loss of control. As observed during the evaluation of these systems, there is a substantial pedagogical gap in which the interface assumes a high level of epistemic engagement from the performer. Without a dedicated period of learning the instrument to understand the machine's specific quirks and latent boundaries, the process of contextual navigation can easily lead to performer frustration. Therefore, the success of this co-creative ecosystem relies heavily on the performer's willingness to relinquish absolute authorial command and adapt to a tool that mimics agency by actively constraining the universe of possibilities.

In practice, this resistance forces the performer to engage in epistemic action (e.g., performing to learn the limits of the space), rather than just pragmatic action (performing to execute a known score). In the *Echoes of the Unseen* case study, material resistance is an aesthetic property that defines the work's formal structure. The performer does not simply execute a score but probes the manifold to discover its density. Performers reported that the system offered varying levels of

resistance depending on the interpolation path chosen via the one-knob interface. For instance, navigating toward low-density regions required more drastic physical gestures, which the performer interpreted as the machine pushing back against unconventional timbre transformations. This creative friction generates a sense of liveness, ensuring the system functions as a dynamic, mutable interface rather than a passive processor. When the performer sees the latent space as a lively, musical entity with its own unique character, it creates a connection between the human and the machine. Since the machine does not simply copy what the performer plays, the performer has to dynamically adjust and adapt. In the end, *Echoes of the Unseen* shows that material resistance is necessary for contextual navigation. Without this creative friction, the system just processes information passively. However, when the performer engages with these challenges, the system turns into a dynamic, mutable interface that helps shape the music in surprising and innovative ways.

13.2.2 Contextual Navigation (SQ2.2)

Finally, evaluating this contextual navigation through performative practice, SQ2.2 questions:

To what extent can the material resistance provided by the machine's encoded contextual knowledge and high-dimensional latent space facilitate a state of contextual navigation, where authorship is redistributed between human and machine?

In the Aethra system, contextual navigation is fostered through a co-creative loop where the performer acts as a navigator. The one-knob interface reduces the dimensionality of the control space, allowing the performer to focus on the qualitative attributes present within their own and the machine's sound. Qualitative feedback from the performer also suggested that, even though the system still exhibited drift (biasing the system toward high-density locations of the perceptual manifold), performers became more inclined to treat it as productive material resistance rather than a technical imperfection. Crucially, this drift serves a vital aesthetic function by constraining the generated material to the specific context of the piece. Because sparsely populated regions of the latent space generally represent sonic states that are stylistically out-of-context or morphologically invalid, the system's natural bias toward dense clusters effectively acts as an encoded topological boundary. This structural constraint prevents the improvisation from fracturing the established musical vocabulary, ensuring that the machine's contextually bounded possibilities remain cohesive and firmly anchored within the authentic sonic identity of the work.

This creates a dynamic interaction that functions strictly through mimicking agency, avoiding the problematic assumption that an AI possesses actual intentionality or volition. Rather than having a true intention, the machine provides a deep contextual knowledge of the work's universe, represented by the topological shape and structural organization of its latent space. This encoded knowledge actively restricts the universe of generative possibilities, allowing the performer, whose navigational gestures represent the genuine human intent, to engage in a highly focused contextual navigation. The resulting music is a heteronomous work, where the authorial voice is entangled.

This challenges state-of-the-art notions of authorship in AI music, typically seen as either strictly human-led or machine-autonomous, suggesting instead a distributed authorship where the work only exists in the moment of navigation through the interface.

In BroadcastJSB, this paradigm is applied to the harmonic structure of J.S. Bach's chorales. The system fundamentally redefines the performer's role, shifting them from a traditional note-player executing a fixed score to a harmonic navigator within a continuous topological space. By doing so, the traditional concept of the closed musical artifact is successfully rethought, as authorship becomes distributed across the interface, the encoded topological boundaries of the latent space, and the genuine directional intent of the human navigator.

Within this framework, material resistance is physically experienced by the performer as tonal momentum. Because the MusicVAE latent space naturally discovers and organizes itself around structural principles like the circle of fifths, a performer attempting to navigate toward distant tonal regions encounters active resistance based on the learned statistical density of the Bach corpus. This establishes a form of harmonic gravity, where specific paths of modulation feel perceptually heavier or more resistant than others.

Contextual navigation is specifically fostered through the implementation of distinct decoder strategies, which function as the machine's contextually bounded possibilities. Within the NEAR mode, the fundamental harmonic identity of a chorale remains stable while localized vertical sonorities vary, whereas the VECTOR mode strictly preserves the tonal momentum and directional trajectory of the musical phrase. By aligning the system's encoded geometric pathways with the performer's directional intent, the strategy transforms aimless free exploration into a highly structured contextual navigation guided by the corpus's encoded harmonic logic.

Simultaneously, BroadcastJSB's radio metaphor functions as a cognitive bridge, encouraging performers to approach latent interpolated noise not as a technological defect, but as a deliberate expressive terrain. By treating the chorales not as a prescriptive score but as a dynamic exploratory space, BroadcastJSB demonstrates that meaningful contextual navigation occurs only when the performer agrees to navigate the topological resistance of the machine as a legitimate, co-creative musical constraint.

Ultimately, this paradigm of contextual navigation directly addresses a critical vulnerability in the ecology of digital preservation: the threat of technological obsolescence reducing interactive works to dead artifacts, such as fixed audio recordings or written scores. By decoupling the offline generation of latent representations from the real-time performance interfaces, BroadcastJSB and Aethra shift the locus of preservation from a fragile software state to the resilient, mathematical structure of the latent space itself. Rather than attempting to archive a highly specific hardware setup that will inevitably decay, these systems preserve the stylistic nature and generative constraints of the musical corpus as serialized data. Consequently, the traditional concept of the closed musical artifact is successfully rethought. By maintaining the contextually bounded possibilities that provide material resistance, the musical work is conserved not as a static execution, but as a living, co-improvisational ecosystem that can be continually migrated to new platforms and navigated by future post-performers.

13.3 Summary

In this discussion, we explored how shifting from a fixed artifact to a mutable interface can fundamentally change our understanding and experience of music creation. Addressing our first research question (RQ1), we discovered that unsupervised latent spaces should not be seen as mere black boxes. The structured environments that they create enable them to detect and preserve official musical theory concepts. The notion of a mutable work maintains its unity through the tonal functional clusters in MusicVAE and the morphological grouping of extended techniques in RAVE. The introduction of intent-aware navigation methods, which include k -NN+AD and motion-sensitive taxonomies, changes music-making from score following into a process that allows artists to investigate three-dimensional music content through their performances.

The practical uses (RQ2) shown in Aethra and BroadcastJSB demonstrate that contextual navigation functions as an active process that requires the performer to continuously interact with the system's topological boundaries. The creative tension needed to transform the machine from a passive tool into a dynamic, mutable interface emerges from the material resistance that exists within neural manifolds.

Furthermore, this research demonstrates that Explainable AI (XAI) serves as the fundamental bridge that connects abstract latent representation (RQ1) to performative agency (RQ2). Aethra and BroadcastJSB use XAI to improve transparency, which allows performers to legibly engage with machine operations while navigating an entangled performative network. Our research findings create a new framework for electroacoustic practice because they depict musical composition as an interactive musical space that performers can explore through their ability to navigate the restricted universe of the work in a post-digital environment. Thus, the interface changes from a hidden black box system into an accessible epistemic tool. That happens because representational, interactional, and conceptual transparency are implemented through radio tuning metaphors, latent space map visualization, and aestheticization of interpolation noise. The framework demonstrates that intelligibility functions as a necessary condition for an exploratory dialogue because meaningful contextual navigation can only emerge when the performer is able to legibly engage with the system's contextually bounded possibilities.

Chapter 14

Conclusion

This thesis has explored the ontological and performative shift of the musical work from a fixed artifact to a mutable interface. By combining music-theoretic frameworks with advanced latent models, our research revealed that intelligent musical systems can do more than just automation. They can be intelligible partners in the creative process, refining the work as a navigable manifold. In this sense, we bridged the gap between traditional *werktreue* principles and the fluid requirements of contemporary electroacoustic practice.

14.1 Original Contribution to Knowledge

The primary contribution of this research lies in the operationalization of the mutable work through a two-fold representational and navigational strategy, moving the state-of-the-art toward a highly interactive and symbiotic model of AI-mediated performance. By synthesizing the formal logic of symbolic tonality (Chapters 5-8) with the sub-symbolic fluidity of timbre and gesture (Chapters 9-12), this work redefines the musical work from a fixed artifact into a navigable, high-dimensional manifold.

Specifically, this paradigm shift is supported by three distinct types of contributions:

- **Theoretical Contributions:**

Ontological Shift of the Musical Work: The research formally redefines the contemporary electroacoustic musical work, transitioning it from a static artifact into a mutable interface that functions as a continuous, topological space of possibilities.

Musical Intelligibility of Deep Learning: The thesis empirically proves that black-box deep learning models are musically intelligible. Unsupervised Variational Autoencoders (VAEs) autonomously discover and organize their latent spaces according to formal structural principles of Western music theory, such as the circle of fifths (Chapters 5 and 6). Similarly, we also prove that timbre latent spaces naturally align with morphological clustering of extended saxophone techniques (Chapter 9).

Geometric Formalization of Musical Motion: The research establishes a theoretical taxonomy of musical motions (parallel, oblique, and contrary) to bridge the gap between abstract neural activations and musical intent, demonstrating that traditional contrapuntal relationships manifest as distinct, quantifiable geometric trajectories within a latent space (Chapter 11).

- **Technological Contributions:**

Development of Co-Creative Systems: The creation of two functional, open-access software architectures: BroadcastJSB, a web-based topological instrument that uses a radio metaphor for real-time navigation of Bach’s harmonic spaces (Chapter 8), and Aethra, a co-creative interface for mixed music that synthesizes multidimensional complexity into a single axis of performative control (Chapter 12).

Intent-Aware Navigation Strategies: The formulation of specialized sampling algorithms, specifically k -nearest neighbors (k -NN) and k -NN with angular displacement (k -NN+AD), that successfully solve the issue of latent dead zone by prioritizing directional tonal momentum and functional proximity over simple linear interpolation (Chapters 7 and 10).

Creation of Specialized Open-Access Corpora: The construction of the SaxTimbre (Chapter 9) and Motiv datasets (Chapter 11), which effectively map the high-dimensional timbre landscape and the expressive physical gestures of the tenor saxophone into navigatable neural coordinates.

- **Artistic Contributions:**

Establishment of Contextual Navigation: The research shifts the performative paradigm from rigidly synchronizing with fixed electroacoustic media to an active contextual navigation model, where the performer uses gestural intuition to dynamically explore a machine’s restricted, high-dimensional map.

Live Co-Improvisational Practice: The successful application of these theoretical and technological frameworks in live performance, materialized in the *Echoes of the Unseen* case study, which demonstrates that fixed media can be entirely replaced by a reactive, AI-mediated improvisational partner.

A Post-Performer Organology: The thesis fundamentally redistributes authorship and musical agency, elevating the technological system from a passive execution tool to an active participant, and transitioning the human musician from a strict interpreter of scores into a navigator of a complex sonic ecosystem.

Ultimately, these contributions establish a framework of post-performance organology where the digital interface functions as an interactive computational system rather than a static instrument. Consequently, the performer’s role shifts from executing predetermined actions to navigating a structured, data-driven musical environment.

14.2 Limitations and Critical Reflection

While the results validate the efficacy of the topological instrument concept, several inherent limitations arise from the empirical data and the architectural constraints of the developed systems. These limitations provide a nuanced perspective on the challenges of reconciling deep learning with real-time music performance.

The current implementation of MusicVAE within BroadcastJSB and RAVE within Aethra relies on pre-trained corpora. As noted in the discussion of the epistemic tool, this prevents on-the-fly latent reorganization. The machine's learned physics is essentially frozen post-training. This creates a temporal lag between the evolution of a performer's style and the model's internal logic. Furthermore, the high dimensionality of these latent spaces requires substantial GPU resources for real-time inference with low latency. In the Aethra trials, the one-knob interface successfully distilled dimensionality, but the underlying system still faced a curse of dimensionality where certain regions of the latent space remained sparse, leading to the material resistance being perceived occasionally as technical instability rather than intentional creative friction.

A significant limitation identified in the Motiv investigation (Chapter 11) concerns the machine's interpretive intelligibility, specifically regarding the HMM-based motion recognition. While the taxonomy of parallel, oblique, and contrary motions provided a gestural superstructure, the model's accuracy was highly sensitive to the temporal windowing of the input. For example, oblique motions are frequently misclassified as parallel when the latent shift is too subtle. This suggests that the interpretive intelligibility of the machine is limited by the resolution of its sensors and the categorical rigidity of the HMM, which inherently struggles to interpret the fuzzy or hybrid gestures common in improvisational contexts.

While the DFT alignment provided a robust ground truth for Western tonal logic in BroadcastJSB, the perceptual filters of the current encodings are culturally and theoretically situated. The high correlation with Fourier Phase Spaces is meaningful within the context of the circle of fifths and equal temperament, but it does not necessarily translate to non-Western tuning systems or microtonal traditions. The research recognizes that by aligning latent spaces with specific theoretical ground truths, we risk reinforcing a computational canon that marginalizes musical structures not easily mapped to DFT phases or grid-based pianorolls.

Furthermore, the evaluation of contextual navigation reveals clear limits in how the system generalizes across different users. Qualitative feedback from the Aethra case study indicated that while professional performers often value the material resistance of the manifold, less experienced performers might find the machine's contextually bounded possibilities to be completely opaque or frustrating. This exposes a profound pedagogical gap in the mutable interface: the system assumes a high level of epistemic engagement from the performer. Without a significant period of learning the instrument, the interaction with the system's encoded constraints can collapse into a sense of lost control. Ultimately, this suggests that meaningful contextual navigation is highly dependent not only on the performer's existing musical expertise but also on their willingness to actively relinquish traditional authorial command to the machine.

14.3 Future Work: Toward a Post-Performer Organology

The findings of this thesis open several pathways for a post-performer organology. A logical extension involves the integration of latent diffusion models to explore how semantic and descriptive prompts can guide the navigation of the manifold, allowing for a multi-modal approach to contextual navigation. Beyond these immediate technical shifts, five broader research trajectories emerge.

A primary future challenge is moving beyond static, pre-trained corpora toward systems that exhibit continual learning. Future iterations of the Aethra system could implement online fine-tuning of the RAVE model, allowing the latent space to reorganize itself in real-time based on the performer's specific improvisational vocabulary. This would transform the material resistance of the machine from a fixed property into a plastic, evolving constraint, fostering a more intimate bio-digital feedback loop where the instrument learns from the performer as much as the performer learns from the instrument.

The concept of the mutable interface could be expanded into multi-agent networked performances. In such a scenario, a global latent space would serve as a shared environment for multiple human performers and algorithmic systems. This would move the research toward collective contextual navigation, where the manifold acts as a communal score. Authorship would be further decentralized as the trajectory of one performer shifts the topological density for others, redefining the ensemble as a distributed network of human navigators and machine constraints exploring a common, sonorous territory in a non-Euclidean space.

As a further expansion of Aethra's stylistic adaptability, we preliminarily investigated the system's potential application within a jazz-based improvisational context, with the finality of testing whether the system can move beyond the contemporary electroacoustic context. To achieve that, this exploratory research pursued four solo instrumental jazz standard recordings (guitar, tenor saxophone, trumpet, and piano). We preliminarily examined (listening test) how the system's distance-based and motion-enhanced navigation can support the characteristic idiomatic patterns of jazz. The initial findings indicated that the results were similar to those obtained with *Lamento*. Just as the hybrid navigation framework successfully preserved the morphological continuity and structural stability of the complex saxophone techniques in *Lamento*, the system effectively maintained the stylistic coherence and structural integrity of the jazz vocabulary. This confirms that the combination of distance-based categorization and trajectory alignment provides a robust foundation for meaningful navigation across diverse musical genres, extending the mutable interface into the highly expressive and syncopated domain of jazz improvisation. To build upon these preliminary results, a more in-depth study should focus on integrating continual learning mechanisms, allowing the latent space to adapt and reorganize itself in real-time to the evolving improvisational vocabulary of the individual jazz performers. Furthermore, this extended research should seek to test these capabilities within multi-agent networked performances, exploring how a collective ensemble of improvising musicians might simultaneously navigate and negotiate a shared, dynamic musical topology.

Finally, future investigation must address the decolonization of generative topologies. While this study utilized the DFT to align latent spaces with Western common-tone music theory, future work should develop open-ended topological filters. These would allow the machine to align itself with non-Western tuning systems, microtonal scales, or noise-based aesthetics without imposing a pre-defined theoretical ground truth. This path leads toward a pluralistic organology, where the mutable interface can host a diverse range of cultural and structural logics, ensuring that the future of co-creative AI remains as fluid and diverse as the musical traditions it seeks to augment.

14.4 Final Remarks

Ultimately, this research suggests that the future of electroacoustic practice lies not in the automation of creativity, but in its amplification through intelligent mediation. By reimagining the musical work as a mutable interface, we move beyond the historical tension between the fixity of the score and the ephemerality of performance.

The transition to topological manifolds as sites of musical representation marks a definitive departure from the machine-as-tool toward the machine-as-environment. In this new paradigm, the work is no longer a static object to be reproduced, but a sonorous territory to be inhabited. The machine provides the material resistance necessary for genuine co-creativity, challenging the performer to negotiate with a non-human agency that possesses its own internal logic, momentum, and structural integrity.

As we move further into the digital age, the role of the composer and performer must evolve from the dictates of absolute control toward the art of navigational intent. By embracing the epistemic tool, we do not lose our authorial voice. Instead, we find it embedded within an entangled performative network, a topology where human intuition and algorithmic constraints converge to discover musical forms that were previously unreachable. The high-dimensional landscapes explored in this thesis are not mere mathematical abstractions, but the new frontiers of a post-performerist organology, where the boundary between the self and the instrument is perpetually fluid, and the music is always in a state of becoming.

Epilogue: The Manifold and the Muse

This research began by questioning how the traditional roles of composer and performer must adapt when integrating deep learning into electroacoustic practice. By examining the dense, high-dimensional architectures of neural networks, we demonstrated that the black box nature of AI models can be restructured into a legible, navigable environment. Rather than acting as an independent creator, the machine functions as a complex topographical map, one that encodes and enforces the structural boundaries of the musical works it has learned.

Shifting from static, fixed-media scores to mutable interfaces represents a profound ontological pivot that changes the nature of musical labor. Engaging with these systems requires performers to abandon the safety of a predetermined path and assume the role of active navigators. Within this framework, the material resistance of the neural manifold provides a necessary, productive friction. This resistance transforms the computer from a deterministic execution tool into a dynamic, mutable environment that structurally guides the performer's exploration.

As we move toward a post-human organology, it is clear that the machine's role is not to replace human intuition, but to establish a restricted universe of possibilities. The neural network serves as a new instrumental material characterized by its non-linear encoded constraints. Interacting with these systems demands a different kind of presence from the performer: a heightened epistemic engagement and a more deliberate physical gesture. The system's contextually bounded possibilities are not errors or deviations. Rather, they are encoded structural responses that necessitate a continuous, exploratory dialogue.

The concept of the mutable work demonstrates that musical identity is not strictly bound to a fixed sequence of notes, but can be preserved within the geometric continuity of a topological manifold. In these high-dimensional landscapes, the performative outcome emerges purely from the entanglement of human gestural intent and algorithmic constraint. Ultimately, the objective of this thesis was not to design an autonomous generative agent, but to build a robust framework for contextual navigation. The latent spaces and interfaces developed here provide the structural foundation for future performers to tune into these environments, continuing to explore the topological boundaries of the musical work.

References

- Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–18, Montreal QC Canada, April 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3174156. URL <https://dl.acm.org/doi/10.1145/3173574.3174156>.
- Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Cailion, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating Music From Text, 2023. URL <https://arxiv.org/abs/2301.11325>.
- Memo Akten, Rebecca Fiebrink, and Mick Grierson. Deep Meditations: Controlled Navigation of Latent Space. In *Proceedings of the 2nd Workshop on Machine Learning for Creativity and Design at the Neural Information Processing Systems (NeurIPS)*, Montreal, Canada, 2020. arXiv. doi: 10.48550/ARXIV.2003.00910. URL <https://arxiv.org/abs/2003.00910>.
- Francisco Almeida, Gilberto Bernardes, and Christof Weiß. Mid-Level Harmonic Audio Features for Musical Style Classification. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 210–217, Bengaluru, India, December 2022. doi: <https://doi.org/10.5281/zenodo.7316632>. URL <https://doi.org/10.5281/zenodo.7316632>.
- Manuchehr Aminian, Eric Kehoe, Xiaofeng Ma, Amy Peterson, and Michael Kirby. Exploring Musical Structure Using Tonnetz Lattice Geometry and LSTMs. In Valeria V. Krzhizhanovskaya, Gábor Závodszy, Michael H. Lees, Jack J. Dongarra, Peter M. A. Sloot, Sérgio Brissos, and João Teixeira, editors, *Computational Science – ICCS 2020*, volume 12138, pages 414–424. Springer International Publishing, Cham, 2020. ISBN 978-3-030-50417-5. doi: 10.1007/978-3-030-50417-5_31. URL https://link.springer.com/10.1007/978-3-030-50417-5_31.
- Emmanuel Amiot. The Torii of Phases. In J. Yust, J. Wild, and J.A. Burgoyne, editors, *Mathematics and Computation in Music: 4th International Conference, MCM 13. Lecture Notes in Computer Science*, volume 7937 of *Lecture Notes in Computer Science*, pages 1–18, Berlin, 2013. Springer, Berlin, Heidelberg. ISBN 978-3-642-39356-3. doi: 10.1007/978-3-642-39357-0_1. URL http://link.springer.com/10.1007/978-3-642-39357-0_1.
- Emmanuel Amiot. *Music Through Fourier Space*. Computational Music Science. Springer International Publishing, Cham, 2016. ISBN 978-3-319-45581-5. doi: 10.1007/978-3-319-45581-5. URL <http://link.springer.com/10.1007/978-3-319-45581-5>.
- Gertrude Elizabeth Margaret Anscombe. *Intention*. Harvard University Press, London, United Kingdom, 2nd edition, October 2020. ISBN 978-0-674-00399-6.

- Luc Ardaillon. *Synthesis and Expressive Transformation of Singing Voice*. PhD thesis, Université Pierre et Marie Curie - Paris VI, Paris, November 2017. URL <https://hal.science/tel-01710926>.
- Daniel Arfib, Jean-Michel Couturier, Loic Kessous, and Vincent Verfaillie. Strategies of Mapping Between Gesture Data and Synthesis Model Parameters Using Perceptual Spaces. *Organised Sound*, 7(2):127–144, August 2002. doi: 10.1017/S1355771802002054. URL https://www.cambridge.org/core/product/identifier/S1355771802002054/type/journal_article.
- Jack Armitage. Gagnavera, August 2023a. URL <https://aimc2023.pubpub.org/pub/ppgwfht1>.
- Jack Armitage. Strengjavera, August 2023b. URL <https://aimc2023.pubpub.org/pub/83k6upv8/release/2>.
- Jack Armitage and Thor Magnusson. Agential Scores: Artificial Life for Emergent, Self-Organising and Entangled Music Notation. In *Proceedings of the International Conference on Technologies for Music Notation and Representation – TENOR’2023*, pages 51 – 61, Boston, Massachusetts, USA, 2023. Northeastern University. ISBN 978-0-6481592-7-8. URL https://iil.is/pdf/2023_tenor_armitage_magnusson_agential_scores.pdf.
- Jack Armitage, Victor Shepardson, Miguel Crozzoli, and Thor Magnusson. Commingling Artificial Life and Interactive Machine Learning in Diffractive Artistic Research Practice. In *The 2024 Conference on Artificial Life*, Online, 2024a. MIT Press. doi: 10.1162/isal_a_00731. URL https://www.mitpressjournals.org/doi/abs/10.1162/isal_a_00731.
- Jack Armitage, Victor Shepardson, and Thor Magnusson. Tölvera: Composing With Basal Agencies. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, pages 282–291, Utrecht, Netherlands, October 2024b. Zenodo. doi: 10.5281/ZENODO.13904854. URL <https://zenodo.org/doi/10.5281/zenodo.13904854>.
- Alejandro Ascarate, Leo Lebrat, Rodrigo Santa Cruz, Clinton Fookes, and Olivier Salvado. Improving the Generation of VAEs With High Dimensional Latent Spaces by the Use of Hyperspherical Coordinates. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Rome, Italy, June 2025. IEEE. ISBN 979-8-3315-1042-8. doi: 10.1109/IJCNN64981.2025.11227651. URL <https://ieeexplore.ieee.org/document/11227651/>.
- Gérard Assayag. Human-Machine Co-Creativity. In Bernard Lubat, Gérard Assayag, and Marc Chemillier, editors, *Artificiel / Cyber-Improvisations. Phonofaune, 2021, Dialogiques d’Uzeste.*, volume 1. HAL, May 2021. URL <https://hal.science/hal-03542986>.
- Gérard Assayag, Camilo Rueda, Mikael Laurson, Carlos Agon, and Olivier Delerue. Computer-Assisted Composition at IRCAM: From PatchWork to OpenMusic. *Computer Music Journal*, 23(3):59–72, September 1999. ISSN 0148-9267, 1531-5169. doi: 10.1162/014892699559896. URL <https://direct.mit.edu/comj/article/23/3/59-72/93413>.
- Gérard Assayag, Laurent Bonnasse-Gahot, and Joakim Borg. Cocreative Interaction: Somax2 and the REACH Project. *Computer Music Journal*, 46(4):7–25, December 2022. ISSN 0148-9267, 1531-5169. doi: 10.1162/comj_a_00662. URL <https://direct.mit.edu/comj/article/46/4/7/119103/Cocreative-Interaction-Somax2-and-the-REACH>.

- Guillaume Ballet, Riccardo Borghesi, Peter Hoffmann, and Fabien Lévy. Studio Online 3.0: An Internet 'Killer Application' for Remote Access to IRCAM Sounds and Processing Tools. In *Journées d'Informatique Musicale*, Paris, France, May 1999. URL <https://hal.science/hal-03112091>.
- Berker Banar, Nick Bryan-Kinns, and Simon Colton. A Tool for Generating Controllable Variations of Musical Themes Using Variational Autoencoders With Latent Space Regularisation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):16401–16403, June 2023. ISSN 2374-3468. doi: 10.1609/aaai.v37i13.27059. URL <https://ojs.aaai.org/index.php/AAAI/article/view/27059>.
- Karen Barad. Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter. *Signs: Journal of Women in Culture and Society*, 28(3):801–831, March 2003. ISSN 0097-9740, 1545-6943. doi: 10.1086/345321. URL <https://www.journals.uchicago.edu/doi/10.1086/345321>.
- Karen Michelle Barad. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press, Durham, 2007. ISBN 978-0-8223-3917-5.
- Scott Barton. Circularity in Rhythmic Representation and Composition. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, pages 505–508, Birmingham, United Kingdom, June 2020. Zenodo. doi: 10.5281/ZENODO.4813501. URL <https://zenodo.org/record/4813501>.
- Steve Benford, Marco Amerotti, Bob Sturm, and Juan Martinez Avila. Negotiating Autonomy and Trust When Performing With an AI Musician. In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, pages 1–10, Austin TX USA, September 2024. ACM. ISBN 979-8-4007-0989-0. doi: 10.1145/3686038.3686040. URL <https://dl.acm.org/doi/10.1145/3686038.3686040>.
- Adam Berenzweig, Daniel P. W. Ellis, and Steve Lawrence. Anchor Space for Classification and Similarity Measurement of Music. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, pages I–29, Baltimore, MD, USA, 2003. IEEE. ISBN 978-0-7803-7965-7. doi: 10.1109/ICME.2003.1220846. URL <http://ieeexplore.ieee.org/document/1220846/>.
- Tony Bergstrom, Karrie Karahalios, and John C. Hart. Isochords: Visualizing Structure in Music. In *Proceedings of Graphics Interface 2007 on - GI '07*, page 297, Montreal, Canada, 2007. ACM Press. ISBN 978-1-56881-337-0. doi: 10.1145/1268517.1268565. URL <http://portal.acm.org/citation.cfm?doid=1268517.1268565>.
- Gilberto Bernardes. Interfacing Sounds: Hierarchical Audio-Content Morphologies for Creative Re-Purposing in Eargram 2.0. In Romain Michon and Franziska Schroeder, editors, *Proceedings of the international conference on new interfaces for musical expression*, pages 537–542, Birmingham, UK, July 2020. Birmingham City University. doi: 10.5281/zenodo.4813176. URL https://www.nime.org/proceedings/2020/nime2020_paper103.pdf.
- Gilberto Bernardes, Carlos Guedes, and Bruce Pennycook. EarGram: An Application for Interactive Exploration of Concatenative Sound Synthesis in Pure Data. In Mitsuko Aramaki, Mathieu Barthet, Richardand Kronland-Martinet, and Sølvi Ystad, editors, *From Sounds to Music and Emotions. CMMR 2012. Lecture Notes in Computer Science, vol 7900*, pages

- 110–129. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-41248-6. doi: 10.1007/978-3-642-41248-6_7. URL http://link.springer.com/10.1007/978-3-642-41248-6_7.
- Gilberto Bernardes, Diogo Cocharro, Marcelo Caetano, Carlos Guedes, and Matthew E. P. Davies. A Multi-Level Tonal Interval Space for Modelling Pitch Relatedness and Musical Consonance. *Journal of New Music Research*, 45(4):281–294, October 2016a. ISSN 0929-8215, 1744-5027. doi: 10.1080/09298215.2016.1182192. URL <https://www.tandfonline.com/doi/full/10.1080/09298215.2016.1182192>.
- Gilberto Bernardes, Diogo Cocharro, Carlos Guedes, and Matthew E. P. Davies. Harmony Generation Driven by a Perceptually Motivated Tonal Interval Space. *Computers in Entertainment*, 14(2):1–21, December 2016b. ISSN 1544-3574. doi: 10.1145/2991145. URL <https://dl.acm.org/doi/10.1145/2991145>.
- Gilberto Bernardes, Matthew E. P. Davies, and Carlos Guedes. A Hierarchical Harmonic Mixing Method. In Mitsuko Aramaki, Matthew E. P. Davies, Richard Kronland-Martinet, and Sølvi Ystad, editors, *Music Technology with Swing: 13th International Symposium, CMMR 2017, Matosinhos, Portugal, September 25-28, 2017, Revised Selected Papers 13*, pages 151–170. Springer, Cham, 2018. ISBN 978-3-030-01692-0. doi: 10.1007/978-3-030-01692-0_11. URL http://link.springer.com/10.1007/978-3-030-01692-0_11.
- Gilberto Bernardes, Nádía Carvalho, and Samuel Pereira. Fluidharmony: Defining an Equal-Tempered and Hierarchical Harmonic Lexicon in the Fourier Space. *Journal of New Music Research*, 51(2-3):142–161, May 2022. ISSN 0929-8215, 1744-5027. doi: 10.1080/09298215.2023.2202641. URL <https://www.tandfonline.com/doi/full/10.1080/09298215.2023.2202641>.
- Gonçalo Bernardo and Gilberto Bernardes. Leveraging Compatibility and Diversity in Computer-Aided Music Mashup Creation. *Personal and Ubiquitous Computing*, 27(5):1793–1809, October 2023. ISSN 1617-4909. doi: 10.1007/s00779-022-01702-z. URL <https://link.springer.com/10.1007/s00779-022-01702-z>.
- Paul Best, Jean Bresson, and Diemo Schwarz. Musical Gesture Recognition Using Machine Learning and Audio Descriptors. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–4, La Rochelle, September 2018. IEEE. ISBN 978-1-5386-7021-7. doi: 10.1109/CBML.2018.8516474. URL <https://ieeexplore.ieee.org/document/8516474/>.
- Frédéric Bevilacqua, Norbert Schnell, Jules François, Éric O. Boyer, Diemo Schwarz, and Baptiste Caramiaux. Designing Action–Sound Metaphors Using Motion Sensing and Descriptor-Based Synthesis of Recorded Sound Materials. In Micheline Lesaffre, Pieter-Jan Maes, and Marc Leman, editors, *The Routledge Companion to Embodied Music Interaction*, pages 391–401. Routledge, New York, 1 edition, September 2017. ISBN 978-1-315-62136-4. doi: 10.4324/9781315621364-43. URL <https://www.taylorfrancis.com/books/9781317219736/chapters/10.4324/9781315621364-43>.
- Louis Bigo and Moreno Andreatta. Topological Structures in Computer-Aided Music Analysis. In David Meredith, editor, *Computational Music Analysis*, pages 57–80. Springer International Publishing, Cham, 2016. ISBN 978-3-319-25931-4. doi: 10.1007/978-3-319-25931-4_3. URL http://link.springer.com/10.1007/978-3-319-25931-4_3.

- Jeff A. Bilmes. A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. *International Computer Science Institute*, 4:126, 1998. URL https://www.cs.cmu.edu/~aarti/Class/10701/readings/gentle_tut_HMM.pdf.
- Tim Blackwell, Oliver Bown, and Michael Young. Live Algorithms: Towards Autonomous Computer Improvisers. In Jon McCormack and Mark d’Inverno, editors, *Computers and Creativity*, pages 147–174. Springer, Berlin, Heidelberg, 1st edition, 2012. ISBN 978-3-642-31727-9. doi: https://doi.org/10.1007/978-3-642-31727-9_6.
- Joakim Borg. Somax 2: A Real-Time Framework for Human-Machine Improvisation. Technical report, Institute for Research and Coordination in Acoustics/Music (IRCAM), Paris, France, December 2019. URL http://repmus.ircam.fr/_media/dyci2/report_somax2_joakim_borg-dec.-2019-v1.pdf.
- Joakim Borg, Gérard Assayag, and Mikhail Malt. Somax 2 a Reactive Multi-Agent Environment for Co-Improvisation. In *Proceedings of 19th Sound and Music Computing Conference (SMC 2022)*, Saint-Étienne, France, June 2022. URL <https://hal.science/hal-04001271>.
- Georgina Born. On Musical Mediation: Ontology, Technology and Creativity. *Twentieth-Century Music*, 2(1):7–36, March 2005. ISSN 1478-5722, 1478-5730. doi: 10.1017/S147857220500023X. URL https://www.cambridge.org/core/product/identifier/S147857220500023X/type/journal_article.
- Michal Borsky, Daryush D. Mehta, Jarrad H. Van Stan, and Jon Gudnason. Modal and Nonmodal Voice Quality Classification Using Acoustic and Electroglottographic Features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2281–2291, December 2017. ISSN 2329-9290, 2329-9304. doi: 10.1109/TASLP.2017.2759002. URL <https://ieeexplore.ieee.org/document/8114356/>.
- Dimitri Bouche, Jérôme Nika, Alex Chechile, and Jean Bresson. Computer-Aided Composition of Musical Processes. *Journal of New Music Research*, 46(1):3–14, September 2016. ISSN 0929-8215, 1744-5027. doi: 10.1080/09298215.2016.1230136. URL <https://www.tandfonline.com/doi/full/10.1080/09298215.2016.1230136>.
- Pierre Boulez and Andrew Gerzso. Computers in Music. *Scientific American*, 258(4):44–50, April 1988. ISSN 0036-8733. doi: 10.1038/scientificamerican0488-44. URL <https://www.scientificamerican.com/article/computers-in-music>.
- Rosi Braidotti. *The Posthuman*. Polity Press, Cambridge, UK ; Malden, MA, USA, 1st edition, 2013. ISBN 978-0-7456-4158-4.
- Jean-Pierre Briot and François Pachet. Deep Learning for Music Generation: Challenges and Directions. *Neural Computing and Applications*, 32(4):981–993, February 2020. ISSN 0941-0643, 1433-3058. doi: 10.1007/s00521-018-3813-6. URL <http://link.springer.com/10.1007/s00521-018-3813-6>.
- Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. *Deep Learning Techniques for Music Generation*. Computational Synthesis and Creative Systems. Springer International Publishing (Springer Cham), Switzerland, 1 edition, 2020. ISBN 978-3-319-70163-9. doi: 10.1007/978-3-319-70163-9. URL <http://link.springer.com/10.1007/978-3-319-70163-9>.

- Nicolas Brochec, Marco Fiorini, Mikhail Malt, and Gérard Assayag. Interactive Music Co-Creation With an Instrumental Technique-Aware System: A Case Study With Flute and SomaX2. In *Proceedings of the 2025 International Computer Music Conference (ICMC)*, Boston (MA), United States, June 2025. URL <https://hal.science/hal-05061669>.
- Nick Bryan-Kinns. Reflections on Explainable AI for the Arts (XAIxArts). *Interactions*, 31(1): 43–47, January 2024. ISSN 1072-5520. doi: 10.1145/3636457. URL <https://dl.acm.org/doi/10.1145/3636457>.
- Nick Bryan-Kinns, Berker Banar, Corey Ford, Courtney N. Reed, Yixiao Zhang, Simon Colton, and Jack Armitage. Exploring XAI for the Arts: Explaining Latent Space in Generative Music. In *Proceedings of the 1st Workshop on eXplainable AI Approaches for Debugging and Diagnosis*, pages 1–14. NeurIPS, December 2021. URL <http://arxiv.org/abs/2308.05496>.
- Nick Bryan-Kinns, Bingyuan Zhang, Songyan Zhao, and Berker Banar. Exploring Variational Auto-Encoder Architectures, Configurations, and Datasets for Generative Music Explainable Ai. *Machine Intelligence Research*, 21(1):29–45, February 2024. ISSN 2731-538X. doi: 10.1007/s11633-023-1457-1. URL <https://link.springer.com/10.1007/s11633-023-1457-1>.
- Marco Buongiorno Nardelli, Garland Culbreth, and Miguel Fuentes. Towards a Measure of Harmonic Complexity in Western Classical Music. *Advances in Complex Systems*, 25(05n06): 2240008, August 2022. ISSN 0219-5259, 1793-6802. doi: 10.1142/S0219525922400082. URL <https://www.worldscientific.com/doi/10.1142/S0219525922400082>.
- Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding Disentangling in β -VAE. In *Proceedings of the 2017 NIPS Workshop on Learning Disentangled Representations*, Long Beach Convention Center, Long Beach, 2018. arXiv. doi: 10.48550/ARXIV.1804.03599. URL <https://arxiv.org/abs/1804.03599>.
- Matthew Burtner. Making Noise: Extended Techniques After Experimentalism. *New Music Box*, 6(71), March 2005. URL <https://newmusicusa.org/nmbx/making-noise-extended-techniques-after-experimentalism/>.
- Marcelo Caetano and Xavier Rodet. Sound Morphing by Feature Interpolation. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–164, Prague, Czech Republic, May 2011. IEEE. ISBN 978-1-4577-0538-0. doi: 10.1109/ICASSP.2011.5946365. URL <http://ieeexplore.ieee.org/document/5946365/>.
- Marcelo Freitas Caetano and Xavier Rodet. Automatic Timbral Morphing of Musical Instrument Sounds by High-Level Descriptors. In *Proceedings of the International Computer Music Conference*, pages 11–21, United States, June 2010. URL <https://hal.science/hal-00604390>.
- Antoine Caillon and Philippe Esling. RAVE: A Variational Autoencoder for Fast and High-Quality Neural Audio Synthesis. *Computing Research Repository (CoRR)*, December 2021. URL <http://arxiv.org/abs/2111.05011>.
- Antoine Caillon, Adrien Bitton, Brice Gatinet, and Philippe Esling. Timbre Latent Space: Exploration and Creative Aspects. In *Proceedings of the 2nd International Conference on Timbre (Timbre 2020)*, Thessaloniki, Greece, September 2020. arXiv. doi: 10.48550/ARXIV.2008.01370. URL <https://arxiv.org/abs/2008.01370>.

- Clifton Callender. Continuous Transformations. *Music Theory Online*, 10(3), September 2004. ISSN 1067-3040. doi: 10.30535/mto.10.3.3. URL <http://www.mtosmt.org/issues/mto.04.10.3/mto.04.10.3.callender.pdf>.
- Clifton Callender, Ian Quinn, and Dmitri Tymoczko. Generalized Voice-Leading Spaces. *Science*, 320(5874):346–348, April 2008. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1153021. URL <https://www.science.org/doi/10.1126/science.1153021>.
- Alberto De Campo. Lose Control, Gain Influence - Concepts for Metacontrol. In *Proceedings of the 11th Sound and Music Computing Conference and 40th International Computer Music Conference (SMC/ICMC2014)*, Athens, Greece, September 2014. Zenodo. doi: 10.5281/ZENODO.850506. URL <https://zenodo.org/record/850506>.
- Carlos E. Cancino-Chacón, Maarten Grachten, Werner Goebel, and Gerhard Widmer. Computational Models of Expressive Music Performance: A Comprehensive and Critical Review. *Frontiers in Digital Humanities*, 5:25, October 2018. ISSN 2297-2668. doi: 10.3389/fdigh.2018.00025. URL <https://www.frontiersin.org/article/10.3389/fdigh.2018.00025/full>.
- Linda Candy. Practice Based Research: A Guide. Ph.D Thesis CCS Report: 2006-V1.0 November, University of Technology Sydney, Sydney, NSW, Australia, 2006. URL <https://www.creativityandcognition.com/wp-content/uploads/2011/04/PBR-Guide-1.1-2006.pdf>.
- Jonathan Cannon. Expectancy-Based Rhythmic Entrainment as Continuous Bayesian Inference. *PLOS Computational Biology*, 17(6):e1009025, June 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1009025. URL <https://dx.plos.org/10.1371/journal.pcbi.1009025>.
- Nadia Carvalho. BroadcastJSB: Latent Space Receiver, January 2026a. URL <https://github.com/NadiaCarvalho/BroadcastJSB>.
- Nádia Carvalho. Aethra, January 2026b. URL <https://github.com/NadiaCarvalho/Aethra>.
- Nádia Carvalho and Gilberto Bernardes. Exploring Latent Spaces of Tonal Music Using Variational Autoencoders. In *The International Conference on AI and Musical Creativity (AIMC)*, November 2023. URL <http://arxiv.org/abs/2311.03621>.
- Nádia Carvalho and Gilberto Bernardes. Exploring Sampling Strategies in Latent Spaces for Music Generation. In *Proceedings of the 21st Sound and Music Computing Conference (SMC)*, pages 74–81, Porto, Portugal, July 2024a. ISBN 978-989-35207-5-8. doi: 10.5281/zenodo.14336029. URL <https://doi.org/10.5281/zenodo.14336029>.
- Nádia Carvalho and Gilberto Bernardes. Fourier (common-Tone) Phase Spaces Are in Tune With Variational Autoencoders’ Latent Space. In Thomas Noll, Mariana Montiel, Francisco Gómez, Omar Costa Hamido, José Luis Besada, and José Oliveira Martins, editors, *Mathematics and Computation in Music: 9th International Conference, MCM 2024, Coimbra, PT, June 18–21, Proceedings. Lecture Notes in Computer Science*, volume 14639, pages 305–316, Springer-Verlag, Berlin, Heidelberg, May 2024b. Springer, Cham. ISBN 978-3-031-60638-0. doi: 10.1007/978-3-031-60638-0_24. URL https://link.springer.com/10.1007/978-3-031-60638-0_24.

- Nádia Carvalho and Jorge Sousa. SaxTimbre: A Tenor Saxophone Extended Techniques Dataset, July 2024.
- Nádia Carvalho and Jorge Sousa. Exploring Timbre Latent Spaces: Motion-Enhanced Sampling for Musical Co-Improvisation, May 2025a. URL <https://doi.org/10.6084/m9.figshare.26490382>.
- Nádia Carvalho and Jorge Sousa. Motiv: A Dataset of Latent Space Representations of Musical Phrase Motions, 2025b. URL <https://doi.org/10.7910/DVN/RWCG4>.
- Nádia Carvalho, Jorge Sousa, Gilberto Bernardes, and Henrique Portovedo. Unveiling the Timbre Landscape: A Layered Analysis of Tenor Saxophone in RAVE Models. In *Proceedings of the 21st Sound and Music Computing Conference (SMC)*, pages 523–530, Porto, Portugal, July 2024. ISBN 978-989-35207-5-8. doi: 10.5281/zenodo.14362599. URL <https://zenodo.org/records/14362599>.
- Nádia Carvalho, Jorge Sousa, Gilberto Bernardes, and Henrique Portovedo. Motiv: A Dataset of Latent Space Representations of Musical Phrase Motions. In *Proceedings of Audio Mostly 2025 (AM25)*, Coimbra, Portugal, July 2025a. doi: <https://doi.org/10.1145/3771594.3771602>.
- Nádia Carvalho, Jorge Sousa, Henrique Portovedo, and Gilberto Bernardes. Exploring Timbre Latent Spaces: Motion-Enhanced Sampling for Musical Co-Improvisation. *International Journal of Performance Arts and Digital Media*, pages 1–27, July 2025b. ISSN 1479-4713. doi: 10.1080/14794713.2025.2536409. URL <https://www.tandfonline.com/doi/full/10.1080/14794713.2025.2536409>.
- Carmine-Emanuele Cella. Orchidea: A Comprehensive Framework for Target-Based Computer-Assisted Dynamic Orchestration. *Journal of New Music Research*, 51(1):40–68, January 2022. ISSN 0929-8215, 1744-5027. doi: 10.1080/09298215.2022.2150650. URL <https://www.tandfonline.com/doi/full/10.1080/09298215.2022.2150650>.
- Carmine Emanuele Cella, Daniele Ghisi, Vincent Lostanlen, Fabien Lévy, Joshua Fineberg, and Yan Maresz. OrchideaSOL: A Dataset of Extended Instrumental Techniques for Computer-Aided Orchestration. *Computing Research Repository (CoRR)*, abs/2007.00763, July 2020. URL <http://arxiv.org/abs/2007.00763>.
- Joel Chadabe. Interactive Composing: An Overview. *Computer Music Journal*, 8(1):22, 1984. ISSN 01489267. doi: 10.2307/3679894. URL <https://www.jstor.org/stable/3679894?origin=crossref>.
- Axel Chemla-Romeu-Santos and Philippe Esling. Challenges in Creative Generative Models for Music: A Divergence Maximization Perspective. In *Proceedings of the 3rd Conference on AI Music Creativity*, pages 1–12. AIMC, November 2022. doi: 10.5281/zenodo.7088272. URL <http://arxiv.org/abs/2211.08856>.
- Elaine Chew. The Spiral Array: An Algorithm for Determining Key Boundaries. In *Music and Artificial Intelligence, Second International Conference, ICMAI 2002, Edinburgh, Scotland, UK, September 12-14, 2002, Proceedings*, pages 18–31, 2002. doi: 10.1007/3-540-45722-4_4. URL https://doi.org/10.1007/3-540-45722-4_4.
- Elaine Chew. *Mathematical and Computational Modeling of Tonality - Theory and Applications.*, volume 204 of *International Series in Operations Research and Management Science*. Springer,

2014. ISBN 978-1-4614-9474-4. doi: 10.1007/978-1-4614-9475-1. URL <https://doi.org/10.1007/978-1-4614-9475-1>.
- Elaine Chew and Yun-Ching Chen. Real-Time Pitch Spelling Using the Spiral Array. *Comput. Music. J.*, 29(2):61–76, 2005. doi: 10.1162/0148926054094378. URL <https://doi.org/10.1162/0148926054094378>.
- Michel Chion. *Guide Des Objects Sonores: Pierre Schaeffer Et La Recherche Musicale*. Bibliothèque de recherche musicale. Buchet/Chastel, Paris, France, nouv. éd edition, 1983. ISBN 978-2-7020-1439-4.
- Ching-Hua Chuan and Dorien Herremans. Modeling Temporal Tonal Relations in Polyphonic Music Through Deep Networks With a Novel Image-Based Representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v32i1.11880. URL <https://ojs.aaai.org/index.php/AAAI/article/view/11880>.
- Eric F. Clarke. Categorical Rhythm Perception: An Ecological Perspective. In A. Gabrielsson, editor, *Action and perception in rhythm and music*, pages 19–33. Royal Swedish Academy of Music, Stockholm, 1987a.
- Eric F. Clarke. Levels of Structure in the Organization of Musical Time. *Contemporary Music Review*, 2(1):211–238, January 1987b. ISSN 0749-4467, 1477-2256. doi: 10.1080/07494468708567059. URL <http://www.tandfonline.com/doi/abs/10.1080/07494468708567059>.
- Diogo Cocharro and Gilberto Bernardes. HarT: Smart Tool for Harmonic Content Transformations. In *Proceedings of the 13th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 25–28, Matosinhos, Portugal, September 2017. URL https://www.researchgate.net/profile/Gilberto-Bernardes/publication/320036019_HarT_Smart_Tool_for_Harmonic_Content_Transformations/links/59ca2eada6fdcc451d57e34e/HarT-Smart-Tool-for-Harmonic-Content-Transformations.pdf.
- Diogo Cocharro, Gilberto Bernardes, Gonçalo Bernardo, and Cláudio Lemos. A Review of Musical Rhythm Representation and (Dis)similarity in Symbolic and Audio Domains. In Luísa Correia Castilho, Rui Dias, and José Francisco Pinho, editors, *Perspectives on Music, Sound and Musicology*, volume 10, pages 189–208. Springer International Publishing, Cham, 2021. ISBN 978-3-030-78451-5. doi: 10.1007/978-3-030-78451-5_10. URL https://link.springer.com/10.1007/978-3-030-78451-5_10.
- Richard Cohn. Neo-Riemannian Operations, Parsimonious Trichords, and Their "Tonnetz" Representations. *Journal of Music Theory*, 41(1):1, 1997. ISSN 00222909. doi: 10.2307/843761. URL <http://links.jstor.org/sici?sici=0022-2909%28199721%2941%3A1%3C1%3ANOPTAT%3E2.0.CO%3B2-M&origin=crossref>.
- Richard Cohn. Introduction to Neo-Riemannian Theory: A Survey and a Historical Perspective. *Journal of Music Theory*, 42(2):167, 1998. ISSN 00222909. doi: 10.2307/843871. URL <http://links.jstor.org/sici?sici=0022-2909%28199823%2942%3A2%3C167%3AITNTAS%3E2.0.CO%3B2-I&origin=crossref>.

- Nick Collins. *Towards Autonomous Agents for Live Computer Music: Realtime Machine Listening and Interactive Music Systems*. Ph.D Thesis, University of Cambridge, Cambridge, Mass., 2007a.
- Nick Collins. Live Electronic Music. In Nick Collins and Julio d’Escrivan, editors, *The Cambridge Companion to Electronic Music*, Cambridge companions to music, pages 38–54. Cambridge University Press, Cambridge, 2007b. ISBN 978-1-139-00176-2 978-0-521-86861-7 978-0-521-68865-9. doi: 10.1017/CCOL9780521868617.
- Nick Collins. The Analysis of Generative Music Programs. *Organised Sound*, 13(3):237–248, December 2008. ISSN 1355-7718, 1469-8153. doi: 10.1017/S1355771808000332. URL https://www.cambridge.org/core/product/identifier/S1355771808000332/type/journal_article.
- Josep M. Comajuncosas, Alex Barrachina, John O’Connell, and Enric Gaus. Nuvolet: 3D Gesture-Driven Collaborative Audio Mosaicing. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, pages 252–255, Oslo, Norway, 2011. Zenodo. doi: 10.5281/zenodo.1177987. URL <https://doi.org/10.5281/zenodo.1177987>.
- Arshia Cont. Antescofo: Anticipatory Synchronization and Control of Interactive Parameters in Computer Music. In *International computer music conference (ICMC)*, pages 33–40, Belfast, Ireland, August 2008. URL <https://inria.hal.science/hal-00694803>.
- David Cope. *Experiments in Musical Intelligence*, volume 12 of *The computer music and digital audio series*. A-R Editions, Inc., Madison, WI, 1996.
- Palle Dahlstedt. Ideas and Tools in Material Space - An Extended Spatial Model of Creativity. In *Computational creativity: An interdisciplinary approach*, Germany, 2009. Schloss Dagstuhl-Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing.
- Roger B. Dannenberg. An On-Line Algorithm for Real-Time Accompaniment. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 193–198, 1984.
- Belur V. Dasarathy. *Nearest Neighbour Norms*, volume 1. IEEE Computer Society Press, U.S., Los Alamitos, CA, December 1991. ISBN 978-0-8186-8930-7.
- David L. Davies and Donald W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, April 1979. ISSN 0162-8828, 2160-9292. doi: 10.1109/TPAMI.1979.4766909. URL <http://ieeexplore.ieee.org/document/4766909/>.
- W. Bas de Haas, Matthias Robine, Pierre Hanna, Remco C. Veltkamp, and Frans Wiering. Comparing Approaches to the Similarity of Musical Chord Sequences. In Sølvi Ystad, Mitsuko Aramaki, Richard Kronland-Martinet, and Kristoffer Jensen, editors, *Exploring Music Contents*, volume 6684, pages 242–258. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-23125-4 978-3-642-23126-1. doi: 10.1007/978-3-642-23126-1_16. URL http://link.springer.com/10.1007/978-3-642-23126-1_16.
- Peter Desain and Henkjan Honing. Does Expressive Timing in Music Performance Scale Proportionally With Tempo? *Psychological Research*, 56(4):285–292, July 1994. ISSN 0340-0727, 1430-2772. doi: 10.1007/BF00419658. URL <http://link.springer.com/10.1007/BF00419658>.

- Peter Desain and Henkjan Honing. The Formation of Rhythmic Categories and Metric Priming. *Perception*, 32(3):341–365, March 2003. ISSN 0301-0066, 1468-4233. doi: 10.1068/p3370. URL <https://journals.sagepub.com/doi/10.1068/p3370>.
- Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A Generative Model for Music, 2020. URL <https://arxiv.org/abs/2005.00341>.
- Agostino Di Scipio. Sound Is the Interface: From Interactive to Ecosystemic Signal Processing. *Organised Sound*, 8(3):269–277, December 2003. ISSN 1355-7718, 1469-8153. doi: 10.1017/S1355771803000244. URL https://www.cambridge.org/core/product/identifier/S1355771803000244/type/journal_article.
- Johnny DiBlasi. Tuning Topological Morphologies: Creative Processes of Natural and Artificial Cognitive Systems. In Craig Vear and Fabrizio Poltronieri, editors, *The Language of Creative AI*, pages 235–257. Springer International Publishing, Cham, 2022. ISBN 978-3-031-10959-1 978-3-031-10960-7. doi: 10.1007/978-3-031-10960-7_8. URL https://link.springer.com/10.1007/978-3-031-10960-7_8.
- Pablo Lopez Dieguez and Von-Wun Soo. Variational Autoencoders for Polyphonic Music Interpolation. In *2020 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 56–61, Taipei, Taiwan, December 2020. IEEE. ISBN 978-1-6654-0380-1. doi: 10.1109/TAAI51410.2020.00019. URL <https://ieeexplore.ieee.org/document/9382444/>.
- Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-Track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. In *Proceedings of the thirty-second AAAI conference on artificial intelligence and thirtieth innovative applications of artificial intelligence conference and eighth AAAI symposium on educational advances in artificial intelligence*, AAAI’18/IAAI’18/EAAI’18, New Orleans, Louisiana, USA, 2018. AAAI Press. ISBN 978-1-57735-800-8. doi: 10.5555/3504035.3504040.
- Tianyang Dong, Lulu Yuan, Yuehui Shang, Yang Ye, and Ling Zhang. Direction-Aware Continuous Moving K-Nearest-Neighbor Query in Road Networks. *ISPRS International Journal of Geo-Information*, 8(9):379, August 2019. ISSN 2220-9964. doi: 10.3390/ijgi8090379. URL <https://www.mdpi.com/2220-9964/8/9/379>.
- Jon Drummond. Understanding Interactive Systems. *Organised Sound*, 14(2):124–133, August 2009. ISSN 1355-7718, 1469-8153. doi: 10.1017/S1355771809000235. URL https://www.cambridge.org/core/product/identifier/S1355771809000235/type/journal_article.
- J. C. Dunn†. Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, 4(1): 95–104, January 1974. ISSN 0022-0280. doi: 10.1080/01969727408546059. URL <https://www.tandfonline.com/doi/full/10.1080/01969727408546059>.
- Mark d’Inverno and Michael Luck. *Understanding Agent Systems*. Springer Series on Agent Technology. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-642-07382-3 978-3-662-10702-7. doi: 10.1007/978-3-662-10702-7. URL <http://link.springer.com/10.1007/978-3-662-10702-7>.

- José Echeveste, Jean-Louis Giavitto, and Arshia Cont. A Dynamic Timed-Language for Computer-Human Musical Interaction. Research Report RR-8422, INRIA, December 2013. URL <https://inria.hal.science/hal-00917469>.
- Umberto Eco. *The Open Work*. Harvard University Press, Cambridge, Mass, 1989. ISBN 978-0-674-63976-8.
- Arne Eigenfeldt. Musebots and I: Collaborating With Creative Systems. In Craig Vear and Fabrizio Poltronieri, editors, *The Language of Creative AI*, pages 207–216. Springer International Publishing, Cham, 2022. ISBN 978-3-031-10959-1 978-3-031-10960-7. doi: 10.1007/978-3-031-10960-7_8. URL https://link.springer.com/10.1007/978-3-031-10960-7_8.
- Herbert Eimert. What Is Electronic Music? *Die Reihe*, 1:1–10, 1957.
- Simon Emmerson. *Living Electronic Music*. Ashgate, Aldershot Burlington, Vt, 2007. ISBN 978-0-7546-5548-0 978-0-7546-5546-6.
- Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural Audio Synthesis of Musical Notes With Wavenet Autoencoders. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 1068–1077, Sydney, NSW, Australia, 2017. JMLR.org. doi: <https://dl.acm.org/doi/10.5555/3305381.3305492>. URL <https://dl.acm.org/doi/10.5555/3305381.3305492>.
- Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. GANSynth: Adversarial Neural Audio Synthesis. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, USA, 2019. arXiv. doi: 10.48550/ARXIV.1902.08710. URL <https://arxiv.org/abs/1902.08710>.
- Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differentiable Digital Signal Processing. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Online, Anywhere, 2020. arXiv. doi: 10.48550/ARXIV.2001.04643. URL <https://arxiv.org/abs/2001.04643>.
- Philippe Esling, Axel Chemla-Romeu-Santos, and Adrien Bitton. Generative Timbre Spaces: Regularizing Variational Auto-Encoders With Perceptual Metrics. In *Proceedings of the 21st International Conference on Digital Audio Effects (DAFx)*, pages 23–27, Aveiro, Portugal, October 2018. arXiv. doi: 10.48550/arXiv.1805.08501. URL <http://arxiv.org/abs/1805.08501>.
- Karlheinz Essl. Algorithmic Composition. In Nick Collins and Julio d’Escrivan, editors, *The Cambridge Companion to Electronic Music*, Cambridge companions to music, pages 107–125. Cambridge University Press, Cambridge, 2007. ISBN 978-1-139-00176-2 978-0-521-86861-7 978-0-521-68865-9. doi: 10.1017/CCOL9780521868617.
- Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases With Noise. In *Proceedings of the second international conference on knowledge discovery and data mining*, KDD’96, pages 226–231, Portland, Oregon, 1996. AAAI Press.

- Leonhard Euler. *Tentamen Novae Theoriae Mysicae Ex Certissimis Harmoniae Principiis Dilvicide Expositae*. Ex Typographia Academiae Scientiarvm, 1739. URL https://archive.org/embed/bub_gb_aekmN1V98GcC.
- Sidney Fels. Designing for Intimacy: Creating New Interfaces for Musical Expression. *Proceedings of the IEEE*, 92(4):672–685, April 2004. ISSN 0018-9219. doi: 10.1109/JPROC.2004.825887. URL <http://ieeexplore.ieee.org/document/1278690/>.
- Sidney Fels, Ashley Gadd, and Axel Mulder. Mapping Transparency Through Metaphor: Towards More Expressive Musical Instruments. *Organised Sound*, 7(2):109–126, August 2002. ISSN 1355-7718, 1469-8153. doi: 10.1017/S1355771802002042. URL https://www.cambridge.org/core/product/identifier/S1355771802002042/type/journal_article.
- Jamie Ferguson and Stephen A. Brewster. Investigating Perceptual Congruence Between Data and Display Dimensions in Sonification. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–9, Montreal QC Canada, April 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3174185. URL <https://dl.acm.org/doi/10.1145/3173574.3174185>.
- Rebecca Fiebrink and Perry R. Cook. The Wekinator: A System for Real-Time, Interactive Machine Learning in Music. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, volume 3, pages 711–713, Utrecht, Netherlands, 2010.
- Rebecca Fiebrink and Laetitia Sonami. Reflections on Eight Years of Instrument Creation With Machine Learning. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, pages 237–242, Birmingham, United Kingdom, June 2020. Zenodo. doi: 10.5281/ZENODO.4813334. URL <https://zenodo.org/record/4813334>.
- Rebecca A. Fiebrink and Baptiste Caramiaux. The Machine Learning Algorithm as Creative Musical Tool. In Alex McLean and Roger T. Dean, editors, *Oxford Handbook on Algorithmic Music*, pages 181–208. Oxford University Press, Oxford, UK, February 2018. ISBN 978-0-19-022701-2. doi: 10.1093/oxfordhb/9780190226992.013.23. URL <https://academic.oup.com/edited-volume/28278/chapter/214418708>.
- Rebecca Anne Fiebrink. *Real-time Human Interaction With Supervised Learning Algorithms for Music Composition and Performance*. Princeton University, 2011. ISBN 978-1-124-49189-9.
- Marco Fiorini and Mikhail Malt. Somax2 – a Distributed Co-Creative System for Human-Machine Co-Improvisation. In *Frontiers in Artificial Intelligence and Applications. Volume 368: HHAI 2023: Augmenting Human Intellect*. IOS Press / Sage Publishing, June 2023. doi: 10.3233/FAIA230106. URL <https://ebooks.iospress.nl/doi/10.3233/FAIA230106>.
- N. I. Fisher and A. J. Lee. Nonparametric Measures of Angular-Angular Association. *Biometrika*, 69(2):315, August 1982. ISSN 00063444. doi: 10.2307/2335405. URL <https://www.jstor.org/stable/2335405?origin=crossref>.
- Jonathan Foote and Shingo Uchihashi. The Beat Spectrum: A New Approach to Rhythm Analysis. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, pages 881–884, Tokyo, Japan, 2001. IEEE. ISBN 978-0-7695-1198-6. doi: 10.1109/ICME.2001.1237863. URL <http://ieeexplore.ieee.org/document/1237863/>.

- Michel Foucault. What Is an Author? In Donald F. Bouchard, editor, *Language, Counter-Memory, Practice*, pages 113–138. Cornell University Press, 1977. ISBN 978-0-8014-9204-4.
- François-Xavier Féron and Guillaume Boutard. Instrumentalists on Solo Works With Live Electronics: Towards a Contemporary Form of Chamber Music? In Friedemann Sallis, Valentina Bertolani, Jan Burle, and Laura Zattra, editors, *Live Electronic Music: Composition, Performance, Study*, Routledge research in music, pages 101–122. Routledge, Abingdon, Oxon ; New York, NY, 2018. ISBN 978-1-138-02260-7.
- Guy E. Garnett. The Aesthetics of Interactive Computer Music. *Computer Music Journal*, 25(1): 21–33, 2001. ISSN 01489267, 15315169. doi: 10.1162/014892601300126089. URL <http://www.jstor.org/stable/3681632>.
- Dafydd Gibbon. The Rhythms of Rhythm. *Journal of the International Phonetic Association*, 53(1):233–265, April 2023. ISSN 0025-1003, 1475-3502. doi: 10.1017/S0025100321000086. URL https://www.cambridge.org/core/product/identifier/S0025100321000086/type/journal_article.
- Jon Gillick, Adam Roberts, Jesse Engel, Douglas Eck, and David Bamman. Learning to Groove With Inverse Sequence Transformations. In *Proceedings of the 36th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research (PMLR 97)*, pages 2269–2279, Long Beach, California, 2019. arXiv. doi: 10.48550/ARXIV.1905.06118. URL <https://arxiv.org/abs/1905.06118>.
- Lydia Goehr. *The Imaginary Museum of Musical Works: An Essay in the Philosophy of Music*. Oxford University Press, 1 edition, October 1994. ISBN 978-0-19-823541-5 978-0-19-159750-3. doi: 10.1093/0198235410.001.0001. URL <https://academic.oup.com/book/10658>.
- Francisco Gomez-Martin, Perouz Taslakian, and Godfried T. Toussaint. Evenness Preserving Operations on Musical Rhythms. In *Proceedings of the 2008 C3S2E conference on - C3S2E '08*, page 121, Montreal, Quebec, Canada, 2008. ACM Press. ISBN 978-1-60558-101-9. doi: 10.1145/1370256.1370275. URL <http://portal.acm.org/citation.cfm?doid=1370256.1370275>.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. ISBN 978-0-262-03561-3.
- Google. Gemini 3 Flash (january 2026 Version) [large Language Model], 2026. URL <https://gemini.google.com/>.
- Google Creative Lab, Manny Tan, and Kyle McDonald. The Infinite Drum Machine, 2016. URL <https://experiments.withgoogle.com/ai/drum-machine>.
- Grammarly. Grammarly (january 2026 Version) [computer Software], 2026. URL <https://www.grammarly.com/>.
- John Granzow. Encounterpoint: The Ungainly Instrument as Co-Performer. In Friedemann Sallis, Valentina Bertolani, Jan Burle, and Laura Zattra, editors, *Live Electronic Music: Composition, Performance, Study*, Routledge research in music. Routledge, Abingdon, Oxon ; New York, NY, 2018. ISBN 978-1-138-02260-7.

- John M. Grey. Multidimensional Perceptual Scaling of Musical Timbres. *The Journal of the Acoustical Society of America*, 61(5):1270–1277, May 1977. ISSN 0001-4966, 1520-8524. doi: 10.1121/1.381428. URL <https://pubs.aip.org/jasa/article/61/5/1270/626725/Multidimensional-perceptual-scaling-of-musical>.
- Zixun Guo, Jaeyong Kang, and Dorien Herremans. A Domain-Knowledge-Inspired Music Embedding Space and a Novel Attention Mechanism for Symbolic Music Modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):5070–5077, June 2023. ISSN 2374-3468. doi: 10.1609/aaai.v37i4.25635. URL <https://ojs.aaai.org/index.php/AAAI/article/view/25635>.
- Daniel Gómez-Marín, Sergi Jordà, and Perfecto Herrera. Drum Rhythm Spaces: From Polyphonic Similarity to Generative Maps. *Journal of New Music Research*, 49(5):438–456, October 2020. ISSN 0929-8215, 1744-5027. doi: 10.1080/09298215.2020.1806887. URL <https://www.tandfonline.com/doi/full/10.1080/09298215.2020.1806887>.
- Robert S. Hatten. *A Theory of Virtual Agency for Western Art Music*. Indiana University Press, October 2018. ISBN 978-0-253-03799-2. doi: 10.2307/j.ctv512trj. URL <http://www.jstor.org/stable/10.2307/j.ctv512trj>.
- Kirsten Hermes, Timothy Brookes, and Christopher Hummersone. The Harmonic Centroid as a Predictor of String Instrument Timbral Clarity. In Audio Engineering Society, editor, *Proceedings of the 140th Convention of Audio Engineering Society*, Paris, France, May 2016. Audio Engineering Society.
- Laurie J. Heyer, Semyon Kruglyak, and Shibu Yooseph. Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome Research*, 9(11):1106–1115, November 1999. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.9.11.1106. URL <http://genome.cshlp.org/lookup/doi/10.1101/gr.9.11.1106>.
- Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning Basic Visual Concepts With a Constrained Variational Framework. In *International conference on learning representations*, 2017. URL <https://openreview.net/forum?id=Sy2fzU9gl>.
- Lejaren Hiller and Leonard M. Isaacson. *Experimental Music: Composition With an Electronic Computer*. Greenwood Press, Westport, Conn, 1979. ISBN 978-0-313-22158-3.
- Fred Hosken. *The Pocket: A Theory of Beats as Domains*. Ph.D Thesis, Northwestern University, Evanston, Illinois, USA, 2021.
- Christophe Hourdin, Gerard Charbonneau, and Tarek Moussa. A Sound-Synthesis Technique Based on Multidimensional Scaling of Spectra. *Computer Music Journal*, 21(2):56, 1997. ISSN 01489267. doi: 10.2307/3681108. URL <https://www.jstor.org/stable/3681108?origin=crossref>.
- Cheng-Zhi Anna Huang, Curtis Hawthorne, Adam Roberts, Monica Dinculescu, James Wexler, Leon Hong, and Jacob Howcroft. The Bach Doodle: Approachable Music Composition With Machine Learning at Scale. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR2019)*, Delft, The Netherlands, 2019a. arXiv. doi: 10.48550/ARXIV.1907.06637. URL <https://arxiv.org/abs/1907.06637>.

- Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinulescu, and Douglas Eck. Music Transformer: Generating Music With Long-Term Structure. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019b. URL <https://arxiv.org/abs/1809.04281>.
- Andy Hunt, Marcelo M. Wanderley, and Ross Kirk. Towards a Model for Instrumental Mapping in Expert Musical Interaction. In *Proceedings of the 2000 International Computer Music Conference*, Berlin, Germany, August 2000. Michigan Publishing. URL <https://hdl.handle.net/2027/spo.bbp2372.2000.173>.
- David Huron. Tone and Voice: A Derivation of the Rules of Voice-Leading From Perceptual Principles. *Music Perception*, 19(1):1–64, September 2001. ISSN 0730-7829. doi: 10.1525/mp.2001.19.1.1. URL <https://online.ucpress.edu/mp/article/19/1/1/62106/Tone-and-Voice-A-Derivation-of-the-Rules-of-Voice>.
- Don Ihde. *Technology and the Lifeworld: From Garden to Earth*. Indiana University Press, Bloomington, 1990. ISBN 978-0-253-11416-7.
- Roman Ingarden, Adam Czerniawski, and Jean G. Harrell. *The Work of Music and the Problem of Its Identity*. University of California Press, Berkeley, 1986. ISBN 978-0-520-05529-2.
- Intelligent Instruments Lab. RAVE-Models (revision Ad15daf), 2023. URL <https://huggingface.co/Intelligent-Instruments-Lab/rave-models>.
- Shulei Ji, Xinyu Yang, and Jing Luo. A Survey on Deep Learning for Symbolic Music Generation: Representations, Algorithms, Evaluations, and Challenges. *ACM Computing Surveys*, 56(1):1–39, August 2023. ISSN 0360-0300. doi: 10.1145/3597493. URL <https://dl.acm.org/doi/10.1145/3597493>.
- Luis O. Jiménez and David A. Landgrebe. Supervised Classification in High-Dimensional Space: Geometrical, Statistical, and Asymptotical Properties of Multivariate Data. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, 28(1):39–54, February 1998. ISSN 10946977. doi: 10.1109/5326.661089. URL <http://ieeexplore.ieee.org/document/661089/>.
- Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition, 2002. ISBN 978-0-387-95442-4. doi: 10.1007/b98835. URL <http://link.springer.com/10.1007/b98835>.
- Daniel Jones, Andrew R. Brown, and Mark d’Inverno. The Extended Composer: Creative Reflection and Extension With Generative Tools. In Jon McCormack and Mark d’Inverno, editors, *Computers and Creativity*, pages 175–203. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-31726-2 978-3-642-31727-9. doi: 10.1007/978-3-642-31727-9_7. URL https://link.springer.com/10.1007/978-3-642-31727-9_7.
- Maximos Kaliakatsos-Papakostas. Evolutionary Music, Deep Learning and Conceptual Blending: Enhancing User Involvement in Generative Music Systems. In Penousal Machado, Juan Romero, and Gary Greenfield, editors, *Artificial Intelligence and the Arts*, pages 109–138. Springer International Publishing, Cham, 2021. ISBN 978-3-030-59474-9 978-3-030-59475-6. doi: 10.1007/978-3-030-59475-6_5. URL https://link.springer.com/10.1007/978-3-030-59475-6_5.

- Thomas Kaplan, Lorenzo Jamone, and Marcus Pearce. Probabilistic Modelling of Microtiming Perception. *Cognition*, 239:105532, October 2023. ISSN 00100277. doi: 10.1016/j.cognition.2023.105532. URL <https://linkinghub.elsevier.com/retrieve/pii/S001002772300166X>.
- Daniel Kientzy. *Saxologie - Du Potentiel Acoustico-Expressif Des 7 Saxophones*. PhD thesis, Université de Paris VIII, Paris, France, 1990. URL <http://www.theses.fr/1991PA080535>.
- Jin Hyun Kim and Uwe Seifert. Embodiment and Agency: Towards an Aesthetics of Interactive Performativity. In *Proceedings of the 4th Sound and Music Computing Conference (SMC 2007)*, pages 233–240, Lefkada, Greece, July 2007. National and Kapodistrian University of Athens. ISBN 978-960-6608-75-9. doi: <https://doi.org/10.5281/zenodo.849443>.
- Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR)*, San Diego, 2014. arXiv. doi: 10.48550/ARXIV.1412.6980. URL <https://arxiv.org/abs/1412.6980>.
- Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference for Learning Representations (ICLR)*, Banff, Canada, 2014. arXiv. doi: 10.48550/ARXIV.1312.6114. URL <https://arxiv.org/abs/1312.6114>.
- Gottfried Michael Koenig. Project 1 and Project 2: A Program for Musical Composition. *Electronic Music Reports*, 1:32–44, 1970.
- Teuvo Kohonen. The Self-Organizing Map. *Proceedings of the IEEE*, 78(9):1464–1480, September 1990. ISSN 00189219. doi: 10.1109/5.58325. URL <http://ieeexplore.ieee.org/document/58325/>.
- Carol L Krumhansl. Why Is Musical Timbre So Hard to Understand? *Structure and Perception of Electroacoustic Sound and Music*, pages 43–53, 1989.
- Carol Lynne Krumhansl. *Cognitive Foundations of Musical Pitch*. Number no. 17 in Oxford psychology series. Oxford University Press, Oxford, 1990. ISBN 978-0-19-989320-1.
- Joseph Bernard Kruskal. Multidimensional Scaling by Optimizing Goodness of Fit to a Non-metric Hypothesis. *Psychometrika*, 29(1):1–27, March 1964. ISSN 0033-3123, 1860-0980. doi: 10.1007/BF02289565. URL https://www.cambridge.org/core/product/identifier/S003331230004206X/type/journal_article.
- Ron Kuivila. Open Sources: Words, Circuits and the Notation-Realization Relation in the Music of David Tudor. *Leonardo Music Journal*, 14:17–23, December 2004. ISSN 0961-1215, 1531-4812. doi: 10.1162/0961121043067424. URL <https://direct.mit.edu/lmj/article/63395>.
- Edward W. Large and Mari Riess Jones. The Dynamics of Attending: How People Track Time-Varying Events. *Psychological Review*, 106(1):119–159, January 1999. ISSN 1939-1471, 0033-295X. doi: 10.1037/0033-295X.106.1.119. URL <https://doi.apa.org/doi/10.1037/0033-295X.106.1.119>.
- Bruno Latour. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, July 2005. ISBN 978-0-19-925604-4 978-1-383-03965-8. doi: 10.1093/oso/9780199256044.001.0001. URL <https://academic.oup.com/book/52349>.

- Mikael Laurson and Jacques Duthen. Patchwork, a Graphic Language in Preform. In *Proceedings of the International Computer Music Conference*, volume 1989, 1989, pages 172–175, 1989. URL <http://hdl.handle.net/2027/spo.bbp2372.1989.042>.
- Marc Leman. *Embodied Music Cognition and Mediation Technology*. The MIT Press, August 2007. ISBN 978-0-262-25655-1. doi: 10.7551/mitpress/7476.001.0001. URL <https://direct.mit.edu/books/book/3170/Embodied-Music-Cognition-and-Mediation-Technology>.
- Nathan Lenssen and Deanna Needell. An Introduction to Fourier Analysis With Applications to Music. *Journal of Humanistic Mathematics*, 4(1):72–91, January 2014. ISSN 21598118. doi: 10.5642/jhummath.201401.05. URL <http://scholarship.claremont.edu/jhm/vol4/iss1/5/>.
- Fred Lerdahl. Tonal Pitch Space. *Music Perception*, 5(3):315–349, April 1988. ISSN 0730-7829. doi: 10.2307/40285402. URL <https://online.ucpress.edu/mp/article/5/3/315/62720/Tonal-Pitch-Space>.
- Fred Lerdahl. *Tonal Pitch Space*. Oxford University Press, New York ; Oxford, 2001. ISBN 978-0-19-987037-0. doi: 10.1093/acprof:oso/9780195178296.001.0001.
- Fred Lerdahl and Ray S. Jackendoff. *A Generative Theory of Tonal Music*. The MIT Press, Cambridge, MA, December 1982. ISBN 978-0-262-12094-4.
- Fred Lerdahl and Carol L. Krumhansl. Modeling Tonal Tension. *Music Perception*, 24(4):329–366, April 2007. ISSN 0730-7829, 1533-8312. doi: 10.1525/mp.2007.24.4.329. URL <https://online.ucpress.edu/mp/article/24/4/329/95267/Modeling-Tonal-Tension>.
- David Lewin. *Generalized Musical Intervals and Transformations*. Oxford University Press, Oxford, 1. publ. as an oxford university press paperback edition, 2011. ISBN 978-0-19-975994-1.
- George E. Lewis. Too Many Notes: Computers, Complexity and Culture in "Voyager". *Leonardo Music Journal*, 10:33–39, 2000. ISSN 09611215, 15314812. URL <http://www.jstor.org/stable/1513376>.
- Christian Limberg and Zhe Zhang. Mapping the Audio Landscape for Innovative Music Sample Generation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1207–1213, Phuket Thailand, May 2024. ACM. ISBN 979-8-4007-0619-6. doi: 10.1145/3652583.3657586. URL <https://dl.acm.org/doi/10.1145/3652583.3657586>.
- Beth Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In *Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR)*, Massachusetts, USA, October 2000. URL http://ismir2000.ismir.net/papers/logan_paper.pdf.
- Hugh Christopher Longuet-Higgins. Perception of Melodies. *Nature*, 263(5579):646–653, October 1976. ISSN 0028-0836, 1476-4687. doi: 10.1038/263646a0. URL <https://www.nature.com/articles/263646a0>.
- Hugh Christopher Longuet-Higgins. *Mental Processes: Studies in Cognitive Science*. Number 1 in Explorations in cognitive science. MIT Press, Cambridge, Mass., 1987. ISBN 978-0-262-12119-4.

- Thor Magnusson. Of Epistemic Tools: Musical Instruments as Cognitive Extensions. *Organised Sound*, 14(2):168–176, August 2009. ISSN 1355-7718, 1469-8153. doi: 10.1017/S1355771809000272. URL https://www.cambridge.org/core/product/identifier/S1355771809000272/type/journal_article.
- Thor Magnusson. *Sonic Writing: Technologies of Material, Symbolic and Signal Inscriptions*. Bloomsbury Academic, New York, NY, 2019. ISBN 978-1-5013-1385-1 978-1-5013-1386-8 978-1-5013-1389-9.
- Mikhail Malt and Benny Sluchin. Realizations & Expansions Performance Cases of an Open Work Using Artificial Intelligence Generative Tools. In *Proceedings of the 2025 International Computer Music Conference (ICMC)*, Boston (MA), United States, June 2025. Northeastern University, New England Conservatory, Berklee College of Music, Boston Conservatory at Berklee et Emerson College. URL <https://hal.science/hal-05070597>.
- Peter Manning. *Electronic and Computer Music*. Oxford University Press, New York, 4th ed edition, 2013. ISBN 978-0-19-974639-2.
- Giorgio Mariani, Irene Tallini, Emilian Postolache, Michele Mancusi, Luca Cosmo, and Emanuele Rodolà. Multi-Source Diffusion Models for Simultaneous Music Generation and Separation. In *The 12th International Conference on Learning Representations (ICLR)*, Vienna, Austria, 2024. URL <https://openreview.net/forum?id=h922Qhkmx1>.
- Max V. Mathews and Franklin Richard Moore. GROOVE — a Program to Compose, Store, and Edit Functions of Time. *Communications of the ACM*, 13(12):715–721, December 1970. ISSN 0001-0782, 1557-7317. doi: 10.1145/362814.362817. URL <https://dl.acm.org/doi/10.1145/362814.362817>.
- Stephen McAdams and Bruno L. Giordano. The Perception of Musical Timbre. In Susan Hallam, Ian Cross, and Michael Thaut, editors, *The Oxford Handbook of Music Psychology*, 2nd edn, volume 1, pages 113–124. Oxford University Press, 2nd edition, 2016. doi: 10.1093/oxfordhb/9780198722946.013.12. URL <https://academic.oup.com/edited-volume/34489/chapter/292611024>.
- Stephen McAdams, James W. Beauchamp, and Suzanna Meneguzzi. Discrimination of Musical Instrument Sounds Resynthesized With Simplified Spectrotemporal Parameters. *The Journal of the Acoustical Society of America*, 105(2):882–897, February 1999. ISSN 0001-4966, 1520-8524. doi: 10.1121/1.426277. URL <https://pubs.aip.org/jasa/article/105/2/882/559057/Discrimination-of-musical-instrument-sounds>.
- Jon McCormack. Creative Ecosystems. In Jon McCormack and Mark d’Inverno, editors, *Computers and Creativity*, pages 39–60. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-31726-2 978-3-642-31727-9. URL https://link.springer.com/10.1007/978-3-642-31727-9_2.
- Jon McCormack and Mark d’Inverno, editors. *Computers and Creativity*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-31726-2 978-3-642-31727-9. URL <https://link.springer.com/10.1007/978-3-642-31727-9>.
- James McDermott. Representation Learning for the Arts: A Case Study Using Variational Autoencoders for Drum Loops. In Penousal Machado, Juan Romero, and Gary Greenfield, editors, *Artificial Intelligence and the Arts*, pages 139–161. Springer International Publishing, Cham,

2021. ISBN 978-3-030-59474-9 978-3-030-59475-6. doi: 10.1007/978-3-030-59475-6_6. URL https://link.springer.com/10.1007/978-3-030-59475-6_6.
- Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, September 2018. ISSN 2475-9066. doi: 10.21105/joss.00861. URL <http://joss.theoj.org/papers/10.21105/joss.00861>.
- Martin McKinney and Jeroen Breebaart. Features for Audio and Music Classification. In Holger H. Hoos and David Bainbridge, editors, *Proceedings of the 4th International Conference on Music Information Retrieval*, pages 1–8, Baltimore, Maryland, USA, October 2003. Johns Hopkins University. ISBN 978-0-9746194-0-8.
- Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. SampleRNN: An Unconditional End-to-End Neural Audio Generation Model. In *5th International Conference on Learning Representations (ICLR)*, volume abs/1612.07837, pages 1696–1706, Toulon, France, February 2017. Curran Associates, Inc. ISBN 978-1-7138-7271-9. URL <http://arxiv.org/abs/1612.07837>.
- Alessandro Ilic Mezza, Massimiliano Zanoni, and Augusto Sarti. A Latent Rhythm Complexity Model for Attribute-Controlled Drum Pattern Generation. *EURASIP Journal on Audio, Speech, and Music Processing*, 2023(1):11, February 2023. ISSN 1687-4722. doi: 10.1186/s13636-022-00267-2. URL <https://asmp-urasipjournals.springeropen.com/articles/10.1186/s13636-022-00267-2>.
- Lu Mi, Tianxing He, Core Francisco Park, Hao Wang, Yue Wang, and Nir Shavit. Revisiting Latent-Space Interpolation Via a Quantitative Evaluation Framework. *Computing Research Repository (CoRR)*, abs/2110.06421:1–11, October 2021. URL <http://arxiv.org/abs/2110.06421>.
- Ryszard S. Michalski, Jaime G. Carbonell, and Tom M. Mitchell, editors. *Machine Learning*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1983. ISBN 978-3-662-12405-5. doi: 10.1007/978-3-662-12405-5. URL <http://link.springer.com/10.1007/978-3-662-12405-5>.
- Eduardo Reck Miranda. *Composing Music With Computers*. Music technology series. Routledge, New York, NY, 1st ed edition, 2014. ISBN 978-0-240-51567-0 978-1-136-12093-0.
- René Mogensen. SomSwarm 1 Computational Creativity Music Improvisation: Human-Computer Co-Creativity. In *Proceedings of the 5th International Conference on Live Interfaces*, pages 21–29, Trondheim, Norway, July 2020. URL <https://www.researchcatalogue.net/view/908792/908793>.
- Bhavya Mor, Sunita Garhwal, and Ajay Kumar. A Systematic Literature Review on Computational Musicology. *Archives of Computational Methods in Engineering*, 27(3):923–937, July 2020. ISSN 1134-3060, 1886-1784. doi: 10.1007/s11831-019-09337-9. URL <http://link.springer.com/10.1007/s11831-019-09337-9>.
- Landon Morrison. Timbre Space: On the Flat History of a Multidimensional Metaphor. *Music & Science*, 7:20592043241268720, January 2024. ISSN 2059-2043, 2059-2043. doi: 10.1177/20592043241268720. URL <https://journals.sagepub.com/doi/10.1177/20592043241268720>.

- Fabian C. Moss, Markus Neuwirth, and Martin Rohrmeier. The Line of Fifths and the Co-Evolution of Tonal Pitch-Classes. *Journal of Mathematics and Music*, 17(2):173–197, May 2023. ISSN 1745-9737, 1745-9745. doi: 10.1080/17459737.2022.2044927. URL <https://www.tandfonline.com/doi/full/10.1080/17459737.2022.2044927>.
- Meinard Müller. Dynamic Time Warping. In *Information Retrieval for Music and Motion*, pages 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-74047-6 978-3-540-74048-3. doi: 10.1007/978-3-540-74048-3_4. URL http://link.springer.com/10.1007/978-3-540-74048-3_4.
- Perry Naseck, Lancelot Blanchard, Madhav Lavakare, Kimaya Lecamwasam, and Joseph A. Paradiso. Physical Manifestation of Generative AI Music Systems for Live Performance. In *Proceedings of the SIGGRAPH Asia 2025 Art Papers*, pages 1–6, Hong Kong, December 2025. ACM. ISBN 979-8-4007-2129-8. doi: 10.1145/3757369.3767613. URL <https://dl.acm.org/doi/10.1145/3757369.3767613>.
- Shahan Nercessian. P-RAVE: Improving Rave Through Pitch Conditioning and More With Application to Singing Voice Conversion. In *Proceedings of the 26th International Conference on Digital Audio Effects (DAFx23)*, Copenhagen, Denmark, 2023.
- Philon Nguyen and Eldad Tsabary. Random Walks on Neo-Riemannian Spaces: Towards Generative Transformations. In *Proceedings of the 3rd Conference on AI Music Creativity (AIMC 2002)*, Online, Anywhere, September 2022. Zenodo. doi: 10.5281/ZENODO.7088389. URL <https://zenodo.org/record/7088389>.
- Gerhard Nierhaus. *Algorithmic Composition: Paradigms of Automated Music Generation*. Springer, Wien New York, 2009. ISBN 978-3-211-75540-2.
- Luc Nijs, Micheline Lesaffre, and Marc Leman. The Musical Instrument as a Natural Extension of the Musician. In *Music and its instruments*, pages 467–484. Editions Delatour France, 2013. URL <https://orbilu.uni.lu/10993/60175>.
- Jérôme Nika. *Guiding Human-computer Music Improvisation: Introducing Authoring and Control With Temporal Scenarios*. Ph.D Thesis, Université Pierre et Marie Curie - Paris VI, Paris, France, May 2016. URL <https://theses.hal.science/tel-01398364>.
- Michael Nyman. *Experimental Music: Cage and Beyond*. Music in the 20th century. Cambridge University Press, Cambridge ; New York, 2nd ed edition, 1999. ISBN 978-0-521-65297-1 978-0-521-65383-1.
- Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A Generative Model for Raw Audio. *arXiv preprint*, abs/1609.03499, September 2016. URL <http://arxiv.org/abs/1609.03499>.
- Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. This Time With Feeling: Learning Expressive Musical Performance. *Neural Computing and Applications*, 32(4):955–967, February 2020. ISSN 0941-0643, 1433-3058. doi: 10.1007/s00521-018-3758-9. URL <http://link.springer.com/10.1007/s00521-018-3758-9>.
- Noelia Ordiz Castaño. *Jesús Villa Rojo: La Lógica Del Discurso*. Number 17 in Colección Música hispana. ICCMU, Madrid, 2008. ISBN 978-84-89457-12-6.

- Noelia Ordiz Castaño. Aproximación a La Figura Del Compositor Jesús Villa Rojo: Historia Viva De La Creación Musical Española. *Musiker*, 18:171–186, 2011.
- Nicola Orio and Diemo Schwarz. Alignment of Monophonic and Polyphonic Music to a Score. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 155–158, Havana, Cuba, 2001.
- Garth Paine. Interactivity, Where to From Here? *Organised Sound*, 7(3):295–304, December 2002. ISSN 1355-7718, 1469-8153. doi: 10.1017/S1355771802003096. URL https://www.cambridge.org/core/product/identifier/S1355771802003096/type/journal_article.
- Sungjin Park. The Work of Art in the Age of Generative AI: Aura, Liberation, and Democratization. *AI & SOCIETY*, 40(3):1807–1816, March 2025. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-024-01948-6. URL <https://link.springer.com/10.1007/s00146-024-01948-6>.
- Richard Parncutt. A Perceptual Model of Pulse Salience and Metrical Accent in Musical Rhythms. *Music Perception*, 11(4):409–464, July 1994. ISSN 0730-7829. doi: 10.2307/40285633. URL <https://online.ucpress.edu/mp/article/11/4/409/46407/A-Perceptual-Model-of-Pulse-Salience-and-Metrical>.
- Ashis Pati and Alexander Lerch. Attribute-Based Regularization of Latent Spaces for Variational Auto-Encoders. *Neural Computing and Applications*, 33(9):4429–4444, May 2021a. ISSN 0941-0643. doi: 10.1007/s00521-020-05270-2. URL <https://link.springer.com/10.1007/s00521-020-05270-2>.
- Ashis Pati and Alexander Lerch. Is Disentanglement Enough? On Latent Representations for Controllable Music Generation. In *Proceedings of 22nd International Society for Music Information Retrieval Conference (ISMIR)*, volume abs/2108.01450, August 2021b. URL <http://arxiv.org/abs/2108.01450>.
- Ashis Pati, Alexander Lerch, and Gaëtan Hadjeres. Learning to Traverse Latent Spaces for Musical Score Inpainting. In Arthur Flexer, Geoffroy Peeters, Julián Urbano, and Anja Volk, editors, *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, pages 343–351, Delft, The Netherlands, July 2019. URL <http://arxiv.org/abs/1907.01164>.
- Geoffroy Peeters. A Large Set of Audio Features for Sound Description (Similarity and Classification) in the Cuidado Project. CUIDADO Ist Project Report 1, IRCAM, Paris, France, April 2004.
- Geoffroy Peeters. Spectral and Temporal Periodicity Representations of Rhythm for the Automatic Classification of Music Audio Signal. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1242–1252, July 2011. ISSN 1558-7916, 1558-7924. doi: 10.1109/TASL.2010.2089452. URL <http://ieeexplore.ieee.org/document/5607297/>.
- Bruce Pennycook. Who Will Turn the Knobs When I Die? *Organised Sound*, 13(3):199–208, December 2008. ISSN 1355-7718, 1469-8153. doi: 10.1017/S1355771808000290. URL https://www.cambridge.org/core/product/identifier/S1355771808000290/type/journal_article.

- Xenia Pestova. Approaches to Notation in Music for Piano and Live Electronics: The Performer's Perspective. In Friedemann Sallis, Valentina Bertolani, Jan Burle, and Laura Zattra, editors, *Live-Electronic Music: Composition, Performance, Study*, Routledge research in music. Routledge, Abingdon, Oxon ; New York, NY, 2018. ISBN 978-1-138-02260-7.
- Andrew Pickering. *The Mangle of Practice: Time, Agency, and Science*. University of Chicago Press, Chicago, 1995. ISBN 978-0-226-66802-4 978-0-226-66803-1.
- Nuno Peixoto de Pinho. *Processos De Criação E Reutilização Musical a Partir Da Obra Do Compositor Jorge Peixinho*. Ph.D Thesis, Universidade Católica Portuguesa, 2016. URL <http://hdl.handle.net/10400.14/29456>.
- Alberto Pinto. A Dynamical Approach to Music Analysis. *Journal of Mathematics and Music*, 6(3):207–219, November 2012. ISSN 1745-9737, 1745-9745. doi: 10.1080/17459737.2012.752891. URL <http://www.tandfonline.com/doi/abs/10.1080/17459737.2012.752891>.
- Manos Plitsis, Kosmas Kritsis, Maximos Kaliakatsos-Papakostas, Aggelos Pikrakis, and Vassilis Katsourous. Towards a Classification and Evaluation of Symbolic Music Encodings for Rnn Music Generation. In *Proceedings of the 1st Joint Conference on AI Music Creativity*, 8. Stockholm, Sweden, Stockholm, Sweden, October 2020. doi: 10.5281/ZENODO.4285410. URL <https://zenodo.org/doi/10.5281/zenodo.4285410>.
- David Poirier-Quinot, Gaetan Parseihian, and Brian F.G. Katz. Comparative Study on the Effect of Parameter Mapping Sonification on Perceived Instabilities, Efficiency, and Accuracy in Real-Time Interactive Exploration of Noisy Data Streams. *Displays*, 47:2–11, April 2017. ISSN 01419382. doi: 10.1016/j.displa.2016.05.001. URL <https://linkinghub.elsevier.com/retrieve/pii/S0141938216300634>.
- Fabrizio Poltronieri. Towards a Symbiotic Future: Art and Creative AI. In Craig Vear and Fabrizio Poltronieri, editors, *The Language of Creative AI*, pages 29–45. Springer International Publishing, Cham, 2022. ISBN 978-3-031-10959-1 978-3-031-10960-7. doi: 10.1007/978-3-031-10960-7_8. URL https://link.springer.com/10.1007/978-3-031-10960-7_8.
- Mathieu Prang and Philippe Esling. Signal-domain Representation of Symbolic Music for Learning Embedding Spaces. In *The 2020 Joint Conference on AI Music Creativity*, volume abs/2109.03454, pages 1–10, Stockholm, Sweden, October 2020. URL <http://arxiv.org/abs/2109.03454>.
- Nicola Privato and Jeremy Muller. Thales at Guthman Competition, March 2024. URL www.youtube.com/live/j4gHsl5CQRs?si=LK82_4tDJJRt5Qp4&t=1097&cbrd=1.
- Ian Quinn. General Equal-Tempered Harmony (Introduction and Part I). *Perspectives of New Music*, 44(2):114–158, 2006. ISSN 2325-7180. doi: 10.1353/pnm.2006.0010. URL <https://muse.jhu.edu/article/795211>.
- Lawrence Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989. ISSN 00189219. doi: 10.1109/5.18626. URL <http://ieeexplore.ieee.org/document/18626/>.

- Richard R. Randall and Bilal Khan. Lerdahl's Tonal Pitch Space Model and Associated Metric Spaces. *Journal of Mathematics and Music*, 4(3):121–131, November 2010. ISSN 1745-9737, 1745-9745. doi: 10.1080/17459737.2010.529654. URL <http://www.tandfonline.com/doi/abs/10.1080/17459737.2010.529654>.
- Christopher Raphael. Aligning Music Audio With Symbolic Scores Using a Hybrid Graphical Model. *Machine Learning*, 65(2-3):389–409, December 2006. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-006-8415-3. URL <http://link.springer.com/10.1007/s10994-006-8415-3>.
- Priyanshu Rawat, Madhvan Bajaj, Satvik Vats, and Vikrant Sharma. A Comprehensive Study Based on MFCC and Spectrogram for Audio Classification. *Journal of Information and Optimization Sciences*, 44(6):1057–1074, 2023. ISSN 02522667, 21690103. doi: 10.47974/JIOS-1431. URL <https://tarupublications.com/doi/10.47974/JIOS-1431>.
- Bruno H. Repp. Relational Invariance of Expressive Microstructure Across Global Tempo Changes in Music Performance: An Exploratory Study. *Psychological Research*, 56(4):269–284, July 1994. ISSN 0340-0727, 1430-2772. doi: 10.1007/BF00419657. URL <http://link.springer.com/10.1007/BF00419657>.
- Hugo Riemann. *Skizze Einer Neuen Methode Der Harmonielehre*. Breitkopf und Härtel, 1880. URL <https://www.digitale-sammlungen.de/view/bsb00110491?page=1>.
- Curtis Roads. *The Computer Music Tutorial*. MIT Press, Cambridge, Mass, 1996. ISBN 978-0-262-18158-7 978-0-262-68082-0.
- Curtis Roads. *Microsound*. the MIT press, Cambridge (Mass.) London, 2004. ISBN 978-0-262-68154-4.
- Curtis Roads. *Composing Electronic Music: A New Aesthetic*. Oxford University Press, New York, 2015. ISBN 978-0-19-023290-0.
- Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume abs/1803.05428, Stockholm, Sweden, 2018. PMLR 80. URL <http://arxiv.org/abs/1803.05428>.
- Robert Rowe. *Interactive Music Systems: Machine Listening and Composing*. MIT Press, Cambridge, Mass, 1993. ISBN 978-0-262-18149-5.
- Robert Rowe. The Aesthetics of Interactive Music Systems. *Contemporary Music Review*, 18(3): 83–87, January 1999. ISSN 0749-4467, 1477-2256. doi: 10.1080/07494469900640361. URL <http://www.tandfonline.com/doi/abs/10.1080/07494469900640361>.
- Robert Rowe. *Machine Musicianship*. The MIT Press, New York, USA, 2003. ISBN 978-0-262-68149-0. URL <https://mitpress.mit.edu/9780262681490/machine-musicianship>.
- Robert Rowe. Split Levels: Symbolic to Sub-Symbolic Interactive Music Systems. *Contemporary Music Review*, 28(1):31–42, February 2009. ISSN 0749-4467, 1477-2256. doi: 10.1080/07494460802664015. URL <http://www.tandfonline.com/doi/abs/10.1080/07494460802664015>.

- Sam T. Roweis and Lawrence K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, December 2000. ISSN 0036-8075. doi: 10.1126/science.290.5500.2323. URL <https://www.science.org/doi/10.1126/science.290.5500.2323>.
- Hanns Holger Rutz. Agency and Algorithms. *Journal of Science and Technology of the Arts*, pages 73–83 Páginas, January 2016. doi: 10.7559/CITARJ.V8I1.223. URL <https://revistas.ucp.pt/index.php/jsta/article/view/7264>.
- Joel Ryan. Some Remarks on Musical Instrument Design at Steim. *Contemporary Music Review*, 6(1):3–17, January 1991. ISSN 0749-4467, 1477-2256. doi: 10.1080/07494469100640021. URL <http://www.tandfonline.com/doi/abs/10.1080/07494469100640021>.
- Joel Ryan. Effort and Expression. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 414–414. International Computer Music Association, 1992.
- Friedemann Sallis, Valentina Bertolani, Jan Burle, and Laura Zattra, editors. *Live Electronic Music: Composition, Performance, Study*. Routledge research in music. Routledge, Abingdon, Oxon ; New York, NY, 2018. ISBN 978-1-138-02260-7.
- Pedro Sarmiento, Adarsh Kumar, Yu-Hua Chen, Cj Carr, Zack Zukowski, and Mathieu Barthet. GTR-CTRL: Instrument and Genre Conditioning for Guitar-Focused Music Generation with Transformers. In Colin Johnson, Nereida Rodríguez-Fernández, and Sérgio M. Rebelo, editors, *Artificial Intelligence in Music, Sound, Art and Design*, volume 13988, pages 260–275. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-29955-1 978-3-031-29956-8. doi: 10.1007/978-3-031-29956-8_17. URL https://link.springer.com/10.1007/978-3-031-29956-8_17.
- Rob Saunders and Petra Gemeinboeck. Creative AI, Embodiment, and Performance. In Craig Vear and Fabrizio Poltronieri, editors, *The Language of Creative AI*, pages 191–205. Springer International Publishing, Cham, 2022. ISBN 978-3-031-10959-1 978-3-031-10960-7. doi: 10.1007/978-3-031-10960-7_8. URL https://link.springer.com/10.1007/978-3-031-10960-7_8.
- Pierre Schaeffer. *Traité Des Objets Musicaux: Essai Interdisciplines*. Pierres vives. Seuil, Paris, nouv. éd edition, 1966. ISBN 978-2-02-002608-6.
- Pierre Schaeffer. *À La Recherche D'une Musique Concrète*. Pierres vives. Seuil, Paris, [nachdr. der ausg.] paris, seuil, 1952 edition, 1998. ISBN 978-2-02-002572-0.
- Margaret Schedel. Electronic Music and the Studio. In Nick Collins and Julio d'Escriván, editors, *The Cambridge Companion to Electronic Music*, Cambridge companions to music, pages 24–37. Cambridge University Press, Cambridge, 2007. ISBN 978-1-139-00176-2 978-0-521-86861-7 978-0-521-68865-9. doi: 10.1017/CCOL9780521868617.
- Bo N. Schenkman and Vijay Kiran Gidla. Detection, Thresholds of Human Echolocation in Static Situations for Distance, Pitch, Loudness and Sharpness. *Applied Acoustics*, 163:107214, June 2020. ISSN 0003682X. doi: 10.1016/j.apacoust.2020.107214. URL <https://linkinghub.elsevier.com/retrieve/pii/S0003682X19303937>.
- Bill Schottstaedt. Machine Tongues XVII: CLM: Music V Meets Common Lisp. *Computer Music Journal*, 18(2):30, 1994. ISSN 01489267. doi: 10.2307/3680441. URL <https://www.jstor.org/stable/3680441?origin=crossref>.

- Franziska Schroeder and Federico Reuben. AI Improvised Music Duo, August 2023. URL <https://aimc2023.pubpub.org/pub/8x9jxz9a/release/3>.
- Hans-Henning Schulze. Categorical Perception of Rhythmic Patterns. *Psychological Research*, 51 (1):10–15, June 1989. ISSN 0340-0727, 1430-2772. doi: 10.1007/BF00309270. URL <http://link.springer.com/10.1007/BF00309270>.
- Diemo Schwarz. The Caterpillar System for Data-Driven Concatenative Sound Synthesis. In *Digital Audio Effects (DAFx)*, pages 135–140, London, UK, October 2003. URL <https://hal.science/hal-01161127v1>.
- Diemo Schwarz. Corpus-Based Concatenative Synthesis. *IEEE Signal Processing Magazine*, 24 (2):92–104, March 2007. ISSN 1053-5888. doi: 10.1109/MSP.2007.323274. URL <http://ieeexplore.ieee.org/document/4117932/>.
- Diemo Schwarz. The Sound Space as Musical Instrument: Playing Corpus-Based Concatenative Synthesis. In *Proceedings of New Interfaces for Musical Expression (NIME)*, pages 250–253, Ann Arbor, Michigan, United States, 2012. doi: 10.13140/2.1.1416.1284. URL <https://www.researchgate.net/doi/10.13140/2.1.1416.1284>.
- Diemo Schwarz and Norbert Schnell. Sound Search by Content-Based Navigation in Large Databases. In *Proceedings of the Sound and Music Computing Conference (SMC)*, pages 1–1, Porto, Portugal, July 2009. URL <https://hal.science/hal-01161241>.
- Diemo Schwarz, Grégory Beller, Bruno Verbrughe, and Sam Britton. Real-Time Corpus-Based Concatenative Synthesis with CataRT. In *Proceedings of the 9 Int. Conference on Digital Audio Effects (DAFx-06)*, pages 279–282, Montreal, Canada, September 2006. URL <https://hal.science/hal-01161358v1>.
- Diemo Schwarz, Sam Britton, Roland Cahen, and Thomas Goepfer. Musical Applications of Real-Time Corpus-Based Concatenative Synthesis. In *International Computer Music Conference (ICMC)*, pages 47–50, Copenhagen, Denmark, October 2007. URL <https://hal.science/hal-01161376>.
- Diemo Schwarz, Axel Roebel, Chungsin Yeh, and Amaury Laburthe. Concatenative Sound Texture Synthesis Methods and Evaluation. In Pavel Rajmic, František Rund, and Jiří Schimmel, editors, *Proceedings of the 19th International Conference on Digital Audio Effects (dafx)*, pages 217–224, Brno, Czech Republic, September 2016. Brno University of Technology, Faculty of Electrical Engineering and Communication / Jiří Schimmel. URL <https://hal.science/hal-01427393>.
- Simon Schwär, Meinard Müller, and Sebastian J. Schlecht. A Variational Y-Autoencoder for Disentangling Gesture and Material of Interaction Sounds. In *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality (AVAR 2022)*, pages 205–214, Redmond, Washington, USA, August 2022. Audio Engineering Society. URL <http://www.aes.org/e-lib/browse.cfm?elib=21853>.
- Edward A. Shanken. Cybernetics and Art: Cultural Convergence in the 1960s. In Bruce Clarke and Linda Dalrymple Henderson, editors, *From Energy to Information*, pages 255–278. Stanford University Press, August 2002. ISBN 978-1-5036-1947-0. doi: 10.1515/9781503619470-021. URL <https://www.degruyter.com/document/doi/10.1515/9781503619470-021/html>.

- Roger N. Shepard. Geometrical Approximations to the Structure of Musical Pitch. *Psychological Review*, 89(4):305–333, 1982. ISSN 1939-1471, 0033-295X. doi: 10.1037/0033-295X.89.4.305. URL <https://doi.apa.org/doi/10.1037/0033-295X.89.4.305>.
- Christopher Small. *Musicking: The Meanings of Performing and Listening*. Music/culture. Wesleyan Univ. Press [u.a.], Middletown, 1998. ISBN 978-0-8195-2257-3.
- Makis Solomos. *From Music to Sound: The Emergence of Sound in 20th- and 21st-century Music*. Routledge research in music. Routledge, Abingdon, Oxon ; New York, NY, 2020. ISBN 978-0-367-19213-6.
- Paul Stapleton and Tom Davis. Ambiguous Devices: Improvisation, Agency, Touch and Feedthrough in Distributed Music Performance. *Organised Sound*, 26(1):52–64, April 2021. ISSN 1355-7718, 1469-8153. doi: 10.1017/S1355771821000054. URL https://www.cambridge.org/core/product/identifier/S1355771821000054/type/journal_article.
- Karlheinz Stockhausen. *Actualia. Die Reihe*, 1:45–51, 1957.
- Bob L. Sturm, Oded Ben-Tal, Úna Monaghan, Nick Collins, Dorien Herremans, Elaine Chew, Gaëtan Hadjeres, Emmanuel Deruty, and François Pachet. Machine Learning Research That Matters for Music Creation: A Case Study. *Journal of New Music Research*, 48(1):36–55, January 2019. ISSN 0929-8215, 1744-5027. doi: 10.1080/09298215.2018.1515233. URL <https://www.tandfonline.com/doi/full/10.1080/09298215.2018.1515233>.
- Keitaro Tanaka, Ryo Nishikimi, Yoshiaki Bando, Kazuyoshi Yoshii, and Shigeo Morishima. Pitch-Timbre Disentanglement of Musical Instrument Sounds Based on VAE-Based Metric Learning. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 111–115, Toronto, ON, Canada, June 2021. IEEE. ISBN 978-1-7281-7605-5. doi: 10.1109/ICASSP39728.2021.9414059. URL <https://ieeexplore.ieee.org/document/9414059/>.
- Kıvanç Tatar and Philippe Pasquier. Masom: A Musical Agent Architecture Based on Self-Organizing Maps, Affective Computing, and Variable Markov Models. In *Proceedings of the 5th International Workshop on Musical Metacreation (MuMe 2017)*, Atlanta, GA, USA, 2017. URL <https://www.metacreation.net/projects/masom>.
- Kıvanç Tatar, Daniel Bisig, and Philippe Pasquier. Latent Timbre Synthesis. *Neural Computing and Applications*, 33(1):67–84, January 2021. ISSN 0941-0643. doi: 10.1007/s00521-020-05424-2. URL <https://link.springer.com/10.1007/s00521-020-05424-2>.
- Heinrich Taube. Common Music: A Music Composition Language in Common Lisp and CLOS. *Computer Music Journal*, 15(2):21, 1991. ISSN 01489267. doi: 10.2307/3680913. URL <https://www.jstor.org/stable/3680913?origin=crossref>.
- William Teixeira, Marco Fiorini, Mikhail Malt, and Gérard Assayag. A Aplicação Do Somax2 Na Eletrônica Em Tempo Real De Chronos IIIC, De Roberto Victório. *Música Hodie*, 24, December 2024. ISSN 1676-3939, 2317-6776. doi: 10.5216/mh.v24.78611. URL <https://revistas.ufg.br/musica/article/view/78611>.

- Notto J. W. Thelle. *Mixed-Initiative Music Making: Collective Agency in Interactive Music Systems*. Ph.D Thesis, Norwegian Academy of Music, Oslo, 2022. URL <https://nmh.no/en/research/publications/mixed-initiative-music-making>.
- Notto J. W. Thelle and Philippe Pasquier. Spire Muse: A Virtual Musical Partner for Creative Brainstorming. In *NIME 2021*, Shanghai, China, June 2021. PubPub. doi: 10.21428/92fbeb44.84c0b364. URL <https://nime.pubpub.org/pub/wcj8sjee>.
- John Thickstun, Zaid Harchaoui, and Sham Kakade. Learning Features of Music From Scratch. In *5th International Conference on Learning Representations (ICLR)*, pages 1378–1391, Toulon, France, April 2017. Curran Associates, Inc. ISBN 978-1-7138-7271-9. URL <http://arxiv.org/abs/1611.09827>.
- Vincent Tiffon. Le Public Comme 'Musiquant': Exemple De L'installation Immersive Et Interactive XY. In Jean-Paul Fourmentraux, editor, *L'Ère Post-Média: Humanités Digitales Et Cultures Numériques*, Collection "Cultures numériques", pages 163–180. Hermann, Paris, 2012. ISBN 978-2-7056-8326-9.
- Vincent Tiffon. Experiencing Music as Strong Works or as Games: The Examination of Learning Processes in the Production and Reception of Live Electronic Music. In Friedemann Sallis, Valentina Bertolani, Jan Burle, and Laura Zattra, editors, *Live-Electronic Music: Composition, Performance, Study*, Routledge research in music. Routledge, Abingdon, Oxon ; New York, NY, 2018. ISBN 978-1-138-02260-7.
- Neil P. McAngus Todd and Guy J. Brown. A Computational Model of Prosody Perception. In *3rd International Conference on Spoken Language Processing (ICSLP 1994)*, pages 127–130. ISCA, September 1994. doi: 10.21437/ICSLP.1994-35. URL https://www.isca-archive.org/icslp_1994/todd94_icslp.html.
- Petri Toiviainen. Optimizing Self-Organizing Timbre Maps: Two Approaches. In Jaime G. Carbonell, Jö Siekmann, G. Goos, J. Hartmanis, J. Van Leeuwen, and Marc Leman, editors, *Music, Gestalt, and Computing*, volume 1317, pages 335–350. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997. ISBN 978-3-540-63526-0 978-3-540-69591-2. doi: 10.1007/BFb0034124. URL <http://link.springer.com/10.1007/BFb0034124>.
- Godfried Toussaint. Computational Geometric Aspects of Rhythm, Melody, and Voice-Leading. *Computational Geometry*, 43(1):2–22, January 2010. ISSN 09257721. doi: 10.1016/j.comgeo.2007.01.003. URL <https://linkinghub.elsevier.com/retrieve/pii/S092577210900042X>.
- Godfried T. Toussaint. *The Geometry of Musical Rhythm: What Makes a "Good" Rhythm Good?* Chapman & Hall/CRC, Boca Raton, second edition edition, 2019. ISBN 978-1-351-24777-1 978-1-351-24775-7 978-1-351-24776-4 978-1-351-24774-0. doi: 10.1201/9781351247771.
- Barry Truax. Computer Music Composition: The Polyphonic Pod System. *Computer*, 11(8): 40–50, August 1978. ISSN 0018-9162, 1558-0814. doi: 10.1109/C-M.1978.218310. URL <https://ieeexplore.ieee.org/document/1647085/>.
- Meliksah Turker, Alara Dirik, and Pinar Yanardag. Midispace: Finding Linear Directions in Latent Space for Music Generation. In *Creativity and Cognition*, pages 420–427, New York, NY, USA, June 2022. ACM. ISBN 978-1-4503-9327-0. doi: 10.1145/3527927.3532790. URL <https://dl.acm.org/doi/10.1145/3527927.3532790>.

- Dmitri Tymoczko. The Geometry of Musical Chords. *Science*, 313(5783):72–74, July 2006. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1126287. URL <https://www.science.org/doi/10.1126/science.1126287>.
- Dmitri Tymoczko. Scale Theory, Serial Theory and Voice Leading. *Music Analysis*, 27(1):1–49, March 2008. ISSN 0262-5245, 1468-2249. doi: 10.1111/j.1468-2249.2008.00257.x. URL <https://onlinelibrary.wiley.com/doi/10.1111/j.1468-2249.2008.00257.x>.
- Dmitri Tymoczko. *A Geometry of Music: Harmony and Counterpoint in the Extended Common Practice*. Oxford studies in music theory. Oxford University Press, New York, 2011. ISBN 978-0-19-533667-2.
- Dmitri Tymoczko. The Generalized Tonnetz. *Journal of Music Theory*, 56(1):1–52, April 2012. ISSN 0022-2909, 1941-7497. doi: 10.1215/00222909-1546958. URL <https://read.dukeupress.edu/journal-of-music-theory/article/56/1/1/14424/The-Generalized-Tonnetz>.
- Dmitri Tymoczko and Jason Yust. Fourier Phase and Pitch-Class Sum. In Mariana Montiel, Francisco Gomez-Martin, and Octavio A. Agustín-Aquino, editors, *Mathematics and Computation in Music*, volume 11502, pages 46–58. Springer International Publishing, Cham, 2019. ISBN 978-3-030-21391-6 978-3-030-21392-3. doi: 10.1007/978-3-030-21392-3_4. URL https://link.springer.com/10.1007/978-3-030-21392-3_4. Series Title: Lecture Notes in Computer Science.
- George Tzanetakis and Perry R. Cook. Musical Genre Classification of Audio Signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, July 2002. ISSN 1063-6676, 1558-2353. doi: 10.1109/TSA.2002.800560. URL <https://ieeexplore.ieee.org/document/1021072/>.
- Talip Ucar. Bridging the ELBO and MMD. *Computing Research Repository (CoRR)*, abs/1910.13181, 2019. doi: 10.48550/ARXIV.1910.13181. URL <http://arxiv.org/abs/1910.13181>.
- Laurens van der Maaten and Geoffrey Hinton. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL <http://jmlr.org/papers/v9/vandermaaten08a.html>.
- Craig Vear. *The Digital Score Musicianship, Creativity and Innovation*. Routledge, Nottingham, UK, April 2019. ISBN 978-1-138-58668-0. URL https://www.routledge.com/The-Digital-Score-Musicianship-Creativity-and-Innovation/Vear/p/book/9781138586680?srsltid=AfmBOopGBDWd-VpcV9uhbn1Xkwsf14W4j490L_Dz6HMycdAEA9VtqA28.
- Craig Vear. Embodied AI and Musicking Robotics. In Craig Vear and Fabrizio Poltronieri, editors, *The Language of Creative AI*, pages 113–138. Springer International Publishing, Cham, 2022a. ISBN 978-3-031-10959-1 978-3-031-10960-7. doi: 10.1007/978-3-031-10960-7_8. URL https://link.springer.com/10.1007/978-3-031-10960-7_8.
- Craig Vear, editor. *The Routledge International Handbook of Practice-Based Research*. Routledge, Abingdon, Oxon ; New York, NY, 2022b. ISBN 978-0-367-34143-5 978-1-032-18220-9.

- Peter-Paul Verbeek. *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Penn State University Press, University Park, PA, 2021. ISBN 978-0-271-03322-8. doi: 10.1515/9780271033228.
- Gabriel Vigliensoni, Louis McCallum, and Rebecca Fiebrink. Creating Latent Spaces for Modern Music Genre Rhythms Using Minimal Training Data. In *Proceedings of 11th International Conference on Computational Creativity (ICCC'20)*, Coimbra, Portugal, September 2020. Zenodo. doi: 10.5281/ZENODO.7415792. URL <https://zenodo.org/record/7415792>.
- Gabriel Vigliensoni, Phoenix Perry, and Rebecca Fiebrink. A Small-Data Mindset for Generative AI Creative Work. In *Proceedings of the Generative AI and Computer Human Interaction (CHI 2022 Workshop) (GenAICHI)*, Online, Anywhere, May 2022. Zenodo. doi: 10.5281/ZENODO.7086327. URL <https://zenodo.org/record/7086327>.
- Jesús Villa-Rojo. *Lamento*, 1989.
- Peter Vuust, Ole A. Heggli, Karl J. Friston, and Morten L. Kringelbach. Music in the Brain. *Nature Reviews Neuroscience*, 23(5):287–305, May 2022. ISSN 1471-003X, 1471-0048. doi: 10.1038/s41583-022-00578-5. URL <https://www.nature.com/articles/s41583-022-00578-5>.
- Cheng-I Wang, Jennifer Hsu, and Shlomo Dubnov. Machine Improvisation With Variable Markov Oracle: Toward Guided and Structured Improvisation. *Computers in Entertainment*, 14(3):1–18, December 2016. ISSN 1544-3574. doi: 10.1145/2905371. URL <https://dl.acm.org/doi/10.1145/2905371>.
- Tao Wang, Junzhe Liu, Cong Jin, Jianguang Li, and Shihao Ma. An Intelligent Music Generation Based on Variational Autoencoder. In *2020 International Conference on Culture-oriented Science & Technology (ICCST)*, pages 394–398, Beijing, China, October 2020a. IEEE. ISBN 978-1-7281-8138-7. doi: 10.1109/ICCST50977.2020.00082. URL <https://ieeexplore.ieee.org/document/9262797/>.
- Ziyu Wang, Yiyi Zhang, Yixiao Zhang, Junyan Jiang, Ruihan Yang, Junbo Zhao, and Gus Xia. PIANOTREE VAE: Structured Representation Learning for Polyphonic Music. In *Proceedings of 21st International Conference on Music Information Retrieval (ISMIR)*, Montreal, Canada, 2020b. arXiv. doi: 10.48550/ARXIV.2008.07118. URL <https://arxiv.org/abs/2008.07118>.
- Simon Waters. Performance Ecosystems: Ecological Approaches to Musical Interaction. In *Proceedings of Electroacoustic Music Studies Network EMS-07*, Leicester, United Kingdom, 2007. DeMontfort University.
- Gottfried Weber. *Versuch Einer Geordneten Theorie Der Tonsezkunst Zum Selbstunterricht, Mit Anmerkungen Für Gelehrtere, Vol. 1: Grammatik Der Tonsezkunst*. Forgotten Books, S.I., 1817. ISBN 978-0-366-66836-6.
- Marcus Weiss and Giorgio Netti. *The Techniques of Saxophone Playing*. Bärenreiter, 2010. ISBN 978-3-7618-2114-5. URL https://www.baerenreiter.com/en/product/BVK02114?srsltid=AfmBOoqusJejQT_G6pGx52MF0HMXFxVv8FWZgFJJ5JVozWQDeKX6AM2z.

- David Wessel and Matthew Wright. Problems and Prospects for Intimate Musical Control of Computers. *Computer Music Journal*, 26(3):11–22, September 2002. ISSN 0148-9267, 1531-5169. doi: 10.1162/014892602320582945. URL <https://direct.mit.edu/comj/article/26/3/11-22/94758>.
- David L. Wessel. Timbre Space as a Musical Control Structure. *Computer Music Journal*, 3(2): 45, June 1979. ISSN 01489267. doi: 10.2307/3680283. URL <https://www.jstor.org/stable/3680283?origin=crossref>.
- Tom White. Sampling Generative Networks, 2016. URL <http://arxiv.org/abs/1609.04468>.
- Julia Wilkins, Prem Seetharaman, Alison Wahl, and Bryan Pardo. Vocalset: A Singing Voice Dataset. In Emilia Gómez, Xiao Hu 0001, Eric Humphrey, and Emmanouil Benetos, editors, *Proceedings of the 19th International Society for Music Information Retrieval Conference, IS-MIR 2018, Paris, France, September 23-27, 2018*, pages 468–474, 2018. ISBN 978-2-9540351-2-3.
- J. Kent Williams. An Interactive, Multimedia Environment for Exploring Tonal Pitch Space. In Michael E. Auer, editor, *Proceedings of the International Conference of "interactive Computer Aided Learning" ICL2007: EPortfolio and Quality in E-Learning*, page 12 pages, Villach, Austria, 2007. Kassel University Press. URL <https://telearn.hal.science/hal-00197232>.
- Todd Winkler. *Composing Interactive Music: Techniques and Ideas Using Max*. The MIT Press, January 2001. ISBN 978-0-262-73139-3. URL <https://mitpress.mit.edu/9780262731393/composing-interactive-music>.
- Lonce Wyse. Audio Spectrogram Representations for Processing With Convolutional Neural Networks. In *Proceedings of the First International Conference on Deep Learning and Music, Anchorage, USA, May 2017*. doi: 10.48550/ARXIV.1706.09559. URL <https://arxiv.org/abs/1706.09559>.
- Anna Xambó and Gerard Roma. Human–Machine Agencies in Live Coding for Music Performance. *Journal of New Music Research*, 53(1-2):33–46, March 2024. ISSN 0929-8215, 1744-5027. doi: 10.1080/09298215.2024.2442355. URL <https://www.tandfonline.com/doi/full/10.1080/09298215.2024.2442355>.
- Iannis Xenakis. *Formalized Music: Thought and Mathematics in Composition*. Pendragon Press, 1992.
- Matthew Yee-King. Latent Spaces: A Creative Approach. In Craig Vear and Fabrizio Poltronieri, editors, *The Language of Creative AI*, pages 137–154. Springer International Publishing, Cham, 2022. ISBN 978-3-031-10959-1 978-3-031-10960-7. doi: 10.1007/978-3-031-10960-7_8. URL https://link.springer.com/10.1007/978-3-031-10960-7_8.
- Yin-Jyun Luo, Kin Wai Cheuk, Tomoyasu Nakano, Masataka Goto, and Dorien Herremans. Unsupervised Disentanglement of Pitch and Timbre for Isolated Musical Instrument Sounds. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, pages 700–707., Montreal, Canada, October 2020. Zenodo. doi: 10.5281/ZENODO.4245532. URL <https://zenodo.org/record/4245532>.

- Li Yingzhen and Stephan Mandt. Disentangled Sequential Autoencoder. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research (PMLR)*, pages 5670–5679, Stockholm, Sweden, July 2018. PMLR. URL <https://proceedings.mlr.press/v80/yingzhen18a.html>.
- Jason Yust. Applications of DFT to the Theory of Twentieth-century Harmony. In Tom Collins, David Meredith, and Anja Volk, editors, *Mathematics and Computation in Music*, volume 9110, pages 207–218. Springer International Publishing, Cham, 2015a. ISBN 978-3-319-20602-8 978-3-319-20603-5. doi: 10.1007/978-3-319-20603-5_22. URL https://link.springer.com/10.1007/978-3-319-20603-5_22.
- Jason Yust. Schubert’s Harmonic Language and Fourier Phase Space. *Journal of Music Theory*, 59(1):121–181, April 2015b. ISSN 0022-2909. doi: 10.1215/00222909-2863409. URL <https://read.dukeupress.edu/journal-of-music-theory/article/59/1/121/14499/Schubert-s-Harmonic-Language-and-Fourier-Phase>.
- Jason Yust. Harmonic Qualities in Debussy’s “Les Sons Et Les Parfums Tournent Dans L’air Du Soir”. *Journal of Mathematics and Music*, 11(2-3):155–173, September 2017. ISSN 1745-9737, 1745-9745. doi: 10.1080/17459737.2018.1450457. URL <https://www.tandfonline.com/doi/full/10.1080/17459737.2018.1450457>.
- Jason Yust. Geometric Generalizations of the Tonnetz and Their Relation to Fourier Phases Spaces. In *Mathematical Music Theory*, pages 253–277. World Scientific, December 2018. doi: 10.1142/9789813235311_0013. URL https://www.worldscientific.com/doi/abs/10.1142/9789813235311_0013.
- Victor Zappi and Kivanç Tatar. Neural Audio Instruments: Epistemological and Phenomenological Perspectives on Musical Embodiment of Deep Learning. *Frontiers in Computer Science*, 7: 1575168, August 2025. ISSN 2624-9898. doi: 10.3389/fcomp.2025.1575168. URL <https://www.frontiersin.org/articles/10.3389/fcomp.2025.1575168/full>.
- Victor Zappi, Dario Mazzanti, Andrea Brogni, and D Caldwell. Concatenative Synthesis Unit Navigation and Dynamic Rearrangement in vrGrains? In *Proceedings of the 9th Sound and Music Computing Conference (SMC)*, pages 403–408, July 2012.
- Laura Zattra. Collaborating on Composition: The Role of the Musical Assistant at Ircam, Ccrma and Csc. In Friedemann Sallis, Valentina Bertolani, Jan Burle, and Laura Zattra, editors, *Live Electronic Music: Composition, Performance, Study*, Routledge research in music. Routledge, Abingdon, Oxon ; New York, NY, 2018a. ISBN 978-1-138-02260-7.
- Laura Zattra. The Musical Assistant. In Friedemann Sallis, Valentina Bertolani, Jan Burle, and Laura Zattra, editors, *Live-Electronic Music: Composition, Performance, Study*, Routledge research in music. Routledge, Abingdon, Oxon ; New York, NY, 2018b. ISBN 978-1-138-02260-7.
- Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, and Ziwei Liu. Large Motion Model for Unified Multi-Modal Motion Generation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024: 18th European Conf. Proceedings, Part XIII. Lecture Notes in Computer Science, vol 15071*, pages 397–421, Milan, Italy, October 2025. Springer-Verlag. doi: 10.1007/978-3-031-72624-8_23. URL https://link.springer.com/10.1007/978-3-031-72624-8_23.