AAPG2025	MusAIc		JCJC
Coordinated by:	Axel MARMORET	48 months	312,125€
Theme E.2 – Artificial intelligence and data science			

MusAIc – Steerable and Interpretable Music Analysis through Hybrid Models

SUMMARY TABLE OF PERSONS INVOLVED IN THE PROJECT

Partner	Name	First name	Current position	Role & responsibilities in the project	Involvement (person.month) throughout the project's total duration
IMT Atlantique	MARMORET	Axel	Associate professor	Scientific coordinator WP lead	38
	FARRUGIA	Nicolas	Professor	Expertise in Audio Signal Processing PhD direction	8
	GRIPON	Vincent	Professor	Expertise in Deep Learning	4
	To be hired		Intern	Participant WP1	6
	To be hired		PhD student	Participant WP1, WP3	36
	To be hired		Post-doc	Participant WP2, WP3	18
Télécom Paris	RICHARD	Gaël	Professor	Expertise in Deep Learning for Music	2
Univ. Mons	GILLIS	Nicolas	Professor	Expertise in Deep Low-Rank Factorization models	2
Multidisciplinary Consortium – artistic partners				Expertise in Music and Artistic Expression	8

CHANGES WITH RESPECT TO THE PRE-PROPOSAL

The title has slightly changed, replacing "efficient" with "steerable".

The budget has increased slightly (from 303,000€ to 312,125€) to account for overheads while remaining within the +7% authorized margin.

CONTENTS

1 PROPOSAL'S CONTEXT, POSITIONING AND OBJECTIVE(S)

1.1 RESEARCH CONTEXT AND OBJECTIVES

This project aims to develop Artificial Intelligence algorithms for music analysis and understanding that are both **steerable** and **interpretable** while maintaining efficiency. We introduce the overall context of the program, including the techniques we aim to explore in Section 1.1.1 and then derive research questions and objectives in Section 1.1.2.

AAPG2025	MusAIc		JCJC
Coordinated by:	Axel MARMORET	48 months	312,125€
Theme E.2 – Artificial intelligence and data science			

1.1.1 General context

Music is a ubiquitous form of art in our lives, shaping emotions, memories, and cultural identities. Yet, despite its presence in virtually all human societies, music remains difficult to fully analyze and understand. One reason for this complexity is that music combines both structured and expressive elements – mathematical regularities in rhythm and harmony coexist with nuances of interpretation and emotion that resist strict formalization. From a scientific perspective, music is a deeply interdisciplinary phenomenon that intertwines acoustic properties, perceptual mechanisms, and cultural influences, making it a challenging subject for computational analysis [1].

With the recent progress of Artificial Intelligence (AI) in many domains – *e.g.*, computer vision, natural language processing, and speech processing – we might expect AI to aid in music understanding. Indeed, state-of-the-art models have shown remarkable performance in tasks such as automatic transcription [2, 3], source separation [4], and music generation [5]. However, while these models achieve high accuracy, they remain difficult to steer and interpret, limiting their usability for a broad range of experts, including musicians, musicologists, sound engineers, and AI researchers.

To address this, we propose developing AI models that are both **steerable** and **interpretable** while maintaining efficiency for real-world musical applications. **Steerability** refers to the ability to actively guide the outputs of a model, either during learning or inference. For instance, steerability implies the possibility of adjusting a particular model for various musical applications. **Interpretability**, on the other hand, refers to the ability to understand, at least partially, the representations learned by the model. An interpretable model allows users to relate its internal structures to concepts from signal processing, music theory, or computational musicology.

Musicians rely on their intuition and training to navigate the complexities of music. Likewise, AI tools for music must do more than generate high-quality outputs – they should enable users to adjust, refine, and direct results to fit artistic, analytical, and scientific needs. Without this flexibility, their practical value remains limited. Some applications would largely benefit from steerability and interpretability, such as:

- Education, where AI can help musicians analyze complex patterns and improve their practice.
- Performance analysis, providing feedback on technique and expression.
- Musicology, where AI models should align with theoretical and historical perspectives on music.

While our focus is on AI tools for music analysis, the methods and models developed in this project may also contribute to the broader field of interpretable AI. In addition, methods developed in the course of this project may apply to other audio domains (*e.g.*, ecoacoustics, speech) and beyond.

Steerability and interpretability are inherently user-dependent, meaning their implementation must align with the expertise and expectations of those interacting with the model. In this proposal, we aim to first design algorithms that are steerable and interpretable by AI experts, ensuring that they integrate constraints motivated by AI and signal processing principles. These constraints will serve as a starting point for gradually extending steerability and interpretability to music professionals, including (but not limited to) musicians, musicologists, and sound engineers. Ultimately, these models should be developed with and for music professionals, enabling them to shape AI outputs in ways that align with their artistic and analytical goals. Achieving this requires a collaborative, multidisciplinary effort, bridging AI research, musicology, and the performing arts.

By prioritizing steerability and interpretability, we aim to create interactive, adaptable, and high-performing AI tools that bridge the gap between computational efficiency and artistic insight, empowering both researchers and practitioners.

Music Information Retrieval – To ground our ambition in concrete applications, we now turn to the field of Music Information Retrieval (MIR), which offers a rich set of challenges for developing and evaluating AI models. MIR is a research field dedicated to analyzing, organizing, and retrieving musical information [1], which we restrict to audio recordings in this proposal. It encompasses a broad range of tasks that address different aspects of music analysis, with varying levels of success. Among the core challenges, one of the most fundamental is Automatic Music Transcription [6], which converts an audio signal into a symbolic representation such as a musical score. Other essential tasks include Music Source Separation [7], which isolates individual instruments or vocals from a mixed audio signal; Chord Recognition [8] and Melody Extraction [9], which identifies harmonic structures over time; Tempo, Beat & Downbeat Tracking [10], which detects rhythmic

elements such as tempo and downbeats; and Music Structure Analysis [11], which segments a piece into meaningful sections like verses and choruses. In this proposal, we will mainly focus on Automatic Music Transcription (AMT), Music Source Separation (MSS), and Music Structure Analysis (MSA), but the impact of our work could extend beyond these tasks.

Music analysis has traditionally followed two main paradigms: structured methods, like "nonnegative low-rank factorizations," which incorporate prior knowledge about musical structures, and data-driven approaches such as "deep learning methods," which learn representations directly from data with minimal assumptions.

Nonnegative low-rank factorization methods – Structured methods are built upon predefined assumptions to describe musical signals based on known physical, perceptual, or theoretical principles. As a particular example, music notes are often assumed to be harmonic in the frequency domain, meaning that their frequency content follows a specific shape with one fundamental frequency and overtones – multiples of the fundamental. In the time domain, structured methods often assume that the different musical events can be decomposed as a combination of canonical events that can accurately represent all events (*e.g.*, representing the signal as a sequence of individual notes). This latter assumption naturally led to the widespread use of nonnegative low-rank factorization methods – in particular, the Nonnegative Matrix Factorization (NMF) – in MIR [12–18] [19, Chap. 8], with applications to tasks such as AMT, MSS, and MSA (among others). These structured methods present some advantages, notably their steerability – structure, constraints, and prior knowledge can be adapted to the particular setting we want to tackle – and their interpretability – the results they produce can often be linked to human-understandable elements, such as instrumental components or rhythmic structures.

However, nonnegative low-rank factorization methods also present significant limitations. Because they rely on predefined models and assumptions, they can impose oversimplified constraints that do not fully capture the complexity of real-world music [7, 19]. In the case of NMF, for instance, the model assumes that the whole piece of music can be constructed as a linear combination of a few frequency components with minimal variability (only one proportional factor per component over the whole range of frequencies), which may be oversimplistic [15]. Some limitations have already been addressed in more elaborate models (*e.g.*, complex NMF [15] or convolutive models [18]), but not exhaustively. Furthermore, the performances of such models are generally way lower than those of the second type of model, data-driven models under deep learning methods.

Deep learning methods – Data-driven approaches, particularly deep learning models, have revolutionized music analysis by leveraging large-scale datasets to learn rich representations without requiring priors about the structure of the solution [20]. Over the past decade, deep learning models (often inspired by advances in computer vision and natural language processing) have become dominant in MIR (*e.g.*, in tasks such as AMT [2, 3], MSS [4, 21], and MSA [22, 23]), for the simple reason that their performances are generally much higher than those of other models (in particular those of structured methods). The recent development of "foundation" models [24, 25] may further enhance the capabilities of AI in music analysis. In computer vision and natural language processing, foundation models have proven to allow for the extraction of generalized, task-agnostic representations, and we can expect that this property transfers to music.

Despite their impressive performance, deep learning models come with critical drawbacks. As of today, such models are difficult to steer and interpret [20]. When possible, guiding their output requires using a new dataset for guiding the model [26, 27] – which may not be available – and/or requires manipulating highly complex mathematical components [28] – a task that is inaccessible to most musicians and non-experts in AI. In addition, deep learning models lack inherent interpretability, meaning that the representations they generate and the decisions they make cannot be easily understood using existing analytical tools [29].

This project, MusAIc (Steerable and Interpretable Music Analysis through Hybrid Models), aims to develop hybrid models offering steerability and interpretability, while maintaining high performance, by bridging the gap between nonnegative low-rank factorization and deep learning methods. These advancements could lead to essential outcomes, such as: (a) helping musicians refine their practice, (b) enabling new forms of artistic interaction with listeners, (c) assisting musicologists in their research, and (d) providing more efficient and adaptable signal processing tools for music, among others.

AAPG2025	MusAIc		JCJC
Coordinated by:	Axel MARMORET	48 months	312,125€
Theme E.2 – Artificial intelligence and data science			

1.1.2 Research objectives and hypotheses

Research objectives – This proposal is expected to address the three following research objectives:

Obj. 1 *Develop interpretable and high-performing AI models for music analysis.*

Obj. 2 *Develop steerable and high-performing AI models for music analysis.*

Obj. 3 *Collaboratively shape AI models to align with the needs of music professionals, ensuring they are interpretable, steerable, and high-performing in real-world contexts.*

While steerability and interpretability are framed as separate objectives (Objectives 1 and 2), they are inherently interconnected. An interpretable model, whose internal representations can be understood, may be easier to steer, while a steerable model, by allowing guided modifications, could be more easily directed toward interpretability. However, this relationship is not guaranteed. For methodological clarity, we treat them as distinct goals while pursuing both.

An essential objective of this project (Objective 3) is to ensure that its outcomes are accessible and beneficial to communities beyond AI research, particularly those that can leverage advancements in AI for music analysis. These include musicians, musicologists, and music enthusiasts, for whom some level of control is critical. To achieve this objective, we plan to actively involve musicians and musicologists in the later stages of the project once the technical developments reach a sufficiently mature stage. Some music professionals have already been identified and have accepted to take on this project, as introduced in Section 2.1.2.

A key challenge in this project is assessing the potential trade-off between steerability, interpretability, and performance. Indeed, a fundamental reason behind the success of deep learning methods is their ability to autonomously learn data representations that optimize performance based on an objective function. This flexibility, however, is closely tied to the vast number of parameters these models possess, making them inherently difficult to audit, interpret, and steer. Imposing constraints to enhance steerability or interpretability could, in turn, restrict this autonomy and impact efficiency.

While we do not anticipate such a limitation, it remains an important risk to investigate. Even if a trade-off proves inevitable, its implications will vary depending on the context of the application. In creative applications, for instance, musicians may prioritize intuitive, steerable AI tools over highly efficient yet opaque systems that offer little insight into their decisions. Understanding this balance will be crucial to ensuring AI models are not just technically proficient but also practically valuable to end users.

Research hypotheses – In this project, we assume that the structured constraints imposed by low-rank factorization methods are highly relevant to music analysis and allow for interpretability and steerability, while the representations learned by deep learning models capture rich musical features explaining (at least partially) their superior performance. The central hypothesis of the MusAIc project is that these complementary advantages can be jointly exploited by leveraging both methods, with the ultimate goal of designing hybrid models that inherently balance structure and expressivity. Building on this premise, the project is structured around three main research directions, each corresponding to a specific hypothesis.

RH1 *Nonnegative low-rank factorization can achieve higher performance by integrating principles inspired by deep learning.*

Nonnegative low-rank factorization methods have long been valued for their ability to provide structured and interpretable representations in MIR [12–18] [19, Chap. 8]. However, these models are parametrized by a small number of matrices or tensors, often limiting their expressivity and ability to capture complex hierarchical structures present in real-world data.

We hypothesize that representing data through a product of multiple matrices and tensors (each nonnegative), as presented in [30] for matrices, results in richer and more expressive representations than using a single matrix. This expectation is supported both empirically and theoretically:

- Empirical evidence: Depth has been a key factor in the success of deep learning models [31], recognized since early breakthroughs such as AlexNet [32]. The fact that deep architectures continue to dominate in practice, despite training difficulties, is itself a strong empirical argument for their effectiveness.

- Theoretical foundations: While shallow networks can serve as universal approximators in theory [33], practical implementations often limit their expressivity. Research has shown that depth can be exponentially more beneficial than the width in some deep learning models [34] and certain tensor decompositions [35], suggesting that deeper models inherently possess a greater capacity for capturing complex structures.

Deep nonnegative low-rank factorization models are expected to extract hierarchical representations, a phenomenon observed in both deep learning [36] and deep NMF [30]. In MIR, this implies that while standard NMF yields a flat decomposition of a spectrogram, deep NMF and tensor variants can reveal structure across layers: lower layers capture fine-grained spectral features (*e.g.*, notes, timbre), while higher layers encode more abstract elements such as motifs, rhythmic patterns, and larger structural cues. This hierarchical organization aligns with the multi-scale structure of music, where atomic elements like individual notes combine to form chords, motifs, and phrases that shape entire musical pieces.

Beyond depth, we propose to explore large-scale learning in nonnegative low-rank factorizations by introducing batch processing and training epochs, which are standard in deep learning but rarely used in nonnegative low-rank factorization (with exceptions in the context of dictionary learning [37]). This second direction should further enhance the efficiency of structured models, enabling them to harness larger datasets, while it is expected to preserve their interpretability.

This research hypothesis will be examined in Work Package 1, presented in Section 1.3.1.

RH2 *Deep learning models may gain interpretability and steerability by integrating structured constraints.*

Deep learning models have achieved remarkable performance in MIR [2–4, 20–25], largely due to their ability to learn expressive representations from data with minimal human intervention. However, their lack of steerability and interpretability [20, 29] makes it difficult for experts to analyze, refine, or control their outputs.

We propose that structured constraints, such as those derived from nonnegative low-rank factorization, can act as inductive biases that encourage models to learn representations that are both interpretable and steerable. Indeed, nonnegative low-rank factorization inherently structures its learned representations [17], and additional constraints such as sparsity, temporal continuity, or prior knowledge about musical theory (such as harmonicity) [7, 19, 38] further refine this structure by enforcing decompositions that align with human-relevant musical components. By embedding similar constraints into deep learning architectures, we can guide their representations toward structured and musically meaningful features while maintaining their capacity to extract high-level abstractions.

Furthermore, we hypothesize that techniques developed for adaptation – such as LoRA [26] – and compression [39, 40] of deep learning models could be extended to improve interpretability and steerability when combined with constraints such as nonnegativity and sparsity. These constraints are known to encourage more interpretable representations and can also support steerability by enabling users to impose musically meaningful priors, allowing for targeted modifications without requiring full retraining. By exploring these directions, we aim to bridge the gap between the expressive power of deep learning and the goal of producing interpretable and steerable representations for MIR.

This research hypothesis will be examined in Work Package 2, presented in Section 1.3.2.

RH3 *Hybrid models, combining structured and data-driven approaches, can achieve, by design, a balance between performance, interpretability, and steerability.*

As introduced earlier, music analysis has traditionally relied on two contrasting approaches: structured methods like nonnegative low-rank factorizations, which offer interpretable representations grounded in explicit priors, and deep learning models, which extract powerful features from data but lack inherent interpretability and steerability. We define **"hybrid" models** as models that integrate both structured and deep learning approaches by design [41], and we expect hybrid models to be natively efficient, interpretable, and steerable.

Rather than treating interpretability and steerability as afterthoughts, we propose that they should be embedded directly into the model's architecture. Hybrid models provide a principled way to achieve this goal by integrating the strengths of both structured and data-driven approaches. This integration is not merely an additive process but a fundamental rethinking of model design.

The development of such models will build on insights from the previous two research hypotheses. However, to avoid a cold start, we will also take inspiration from recent hybrid approaches [42,43] and leverage neural audio codecs [44, 45]. Neural audio codecs have demonstrated the ability to extract highly efficient representations of audio signals by imposing structured constraints during compression. These models learn compact latent spaces that enable high-quality perceptual reconstruction while preserving the essential structure and detail of the original signal. While their primary objective has been efficient signal reconstruction, their structured latent spaces offer a promising foundation for hybrid models. By modifying their constraints – introducing nonnegativity, sparsity, or explicit decompositions of harmonic and rhythmic structures –, we propose that they can be adapted to enhance interpretability and steerability while maintaining their efficiency.

This approach challenges the dominant performance *vs.* steerability/interpretability paradigm by treating them as compatible design goals. Hybrid models, developed from this perspective, offer a promising direction for advancing AI in music analysis by integrating these objectives from the ground up. More than just a technical solution, we expect that hybrid models will also represent the most promising approach for steerable and interpretable models tailored to music professionals – *e.g.*, musicians, musicologists, sound engineers – by being designed according to their specific needs.

This research hypothesis will be examined in Work Package 3, presented in Section 1.3.3.

Work Package 4, presented in Section 1.3.4, will specifically tackle the collaborative development of efficient, steerable, and interpretable AI models for and with music professionals.

Research plan – Building on our hypotheses, we propose to explore research objectives 1 and 2 in a two-step approach. First, we will independently investigate nonnegative low-rank factorization (WP1) and deep learning models (WP2), as these fields have evolved largely in parallel with limited cross-integration. On the structured side, our objective is to develop deep low-rank factorization models by integrating core ideas from deep learning, such as architectural depth, training routines, and large-scale learning. Conversely, on the deep learning side, we will introduce structured constraints—such as low-rankness, non-negativity, and sparsity – to improve steerability and interpretability. While these separate investigations are necessary, they are not sufficient on their own. Consequently, they will be complemented by the development of hybrid models (WP3) that integrate insights from both structured and deep learning approaches. These models will form a unified paradigm – deep-learning nonnegative low-rank factorization models – designed to combine expressivity, efficiency, and interpretability by construction. Finally, beginning at the midpoint of the project, we will engage in a collaborative refinement of these models with and for music professionals (WP4).

1.2 Position of the project as it relates to the state of the art

1.2.1 Signal processing fundamentals of MIR

MIR aims to analyze sound waves from musical instruments recorded by microphones. It involves understanding both the sources and the alterations introduced by recording conditions [7].

Music sources – Music consists of structured sound waves produced by instruments, each with unique acoustic properties. In Western music, the tonal system is based on twelve notes, each associated with a fundamental frequency (f_0), the primary component of pitch perception. Notes form harmony when arranged into chords, while their timing defines rhythm.

Instruments are categorized as pitched or percussive. Pitched instruments (*e.g.*, strings, wind) generate harmonic spectra, where each overtone follows $f_k = k \times f_0$. The distribution of energy across overtones defines timbre, a key factor in distinguishing instruments. Harmonicity is often used as a constraint in MIR models, either explicitly [38] or learned through data-driven methods [46][(*)]. Percussive instruments (e.g., drums, cymbals) produce transient, inharmonic sounds. This distinction is crucial for music analysis, as pitched instruments primarily contribute to harmony, while percussive instruments define rhythmic elements. For instance, melody estimation relies more on harmonic content [9].

[(*)] We find it interesting to notice that these two approaches reflect the structured *vs.* data-driven duality explored in this proposal.

(a) Raw audio (b) Spectrogram

Figure 1: A waveform and a spectrogram computed from this waveform

Recording conditions – Once emitted, music signals are altered by environmental factors such as reverberation, noise, and echoes, typically modeled by the "room impulse response" [7]. Additionally, sound engineers modify signals in the "mixing process", further shaping the final recording [47].

Waveform and spectrograms – Music signals are naturally expressed as waveforms, as illustrated in Figure 1a, but their complexity – 44,100 digital samples per second in CD quality – makes direct processing computationally demanding and noise-sensitive. Although deep learning can now operate on waveforms [4, 24, 25, 44, 48–51] – even if these methods are still considered suboptimal [52] –, spectrograms remain widely used [2, 21, 23], sometimes in addition to the waveform [4, 24, 25, 44]. A spectrogram, as represented in Figure 1b, is a time-frequency representation derived via Fourier analysis or wavelet transforms, with common forms including Short-Time Fourier Transform (STFT), Mel spectrograms, and MFCCs. Most applications rely on the magnitude or squared magnitude of spectrograms, interpreted as nonnegative matrices.

In this proposal, we primarily focus on applications using spectrogram representations. However, if advancements in waveform-based deep learning methods lead to significant performance improvements, we may consider integrating them as front-end processing techniques.

1.2.2 Nonnegative Low-rank factorization

Definition – Low-rank factorization methods aim to approximate a high-dimensional matrix or tensor by decomposing it into a product of lower-dimensional factors, each constrained to be of low rank compared to the original set of data. Fundamentally, low-rank factorization methods are dimensionality reduction methods, preserving essential structure while reducing redundancy. When constrained to nonnegativity, the original data and factorization factors are all constrained to be parametrized by nonnegative values. In practice, these factorizations yield interpretable, additive ("part-based") decompositions because of the nature of nonnegativity.

We will assume throughout this section that nonnegative low-rank factorization methods are applied to nonnegative matrices X or tensors $\mathcal{X}$, representing the music signal as the (squared) magnitude of a spectrogram.

Nonnegative Matrix Factorization (NMF) – NMF decomposes a nonnegative matrix $X \in \mathbb{R}_+^{m\times n}$ into two nonnegative matrices $W \in \mathbb{R}_+^{m\times k}, H \in \mathbb{R}_+^{k\times n}$, of maximal rank $k \ll \min(m,n)$, such that $X \approx WH$ [53]. In music analysis, W contains basis elements (frequency templates), and H contains activation coefficients (temporal activations), as presented in Figure 2.

NMF is defined as the following optimization problem:

$$\underset{W,H\geq 0}{\arg\min}\, d(X|WH), \tag{1}$$

subject to a loss function $d()$. The choice of the loss function is crucial, as it defines how the reconstruction error is measured and can significantly impact the interpretability and application of the model [54]. In audio processing, it is standard to use either the Frobenius norm, the Kullback-Leibler (KL-) divergence, or the Itakura-Saïto (IS-) divergence [13, 54].

NMF leads to a non-convex, NP-hard optimization problem [55], making exact solutions generally intractable. A common workaround is to iteratively optimize W and H [13, 53–55]. Extensions like Convolutive NMF [18] model time-varying frequency templates, improving performance at higher computational cost.

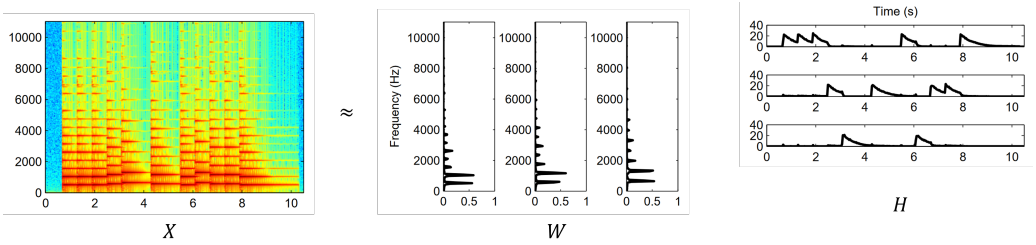


Figure 2: A simplistic example of NMF on the song *"Au Clair de la Lune"*, with $k = 3$. Illustration adapted from [56]. It results in a matrix W with three columns, each corresponding to the frequency template of one note. In H, each line represents their corresponding time activations.

Nonnegative Tensor Factorizations – The matrix model can be extended to tensors – *i.e.* multi-dimensional arrays – allowing for the capture of dependencies across multiple dimensions. This extension enables richer and more structured representations of the data. Two prominent nonnegative tensor factorizations [57] are:

- *Nonnegative Canonical Polyadic (NCP) Decomposition* Also known as nonnegative PARAFAC decomposition, it generalizes NMF to tensors, as a sum of rank-one factors. For a third order tensor $\mathcal{X} \in \mathbb{R}_+^{I \times J \times K}$ with r elements, NPC is defined as:
$$\mathcal{X} \approx \hat{\mathcal{X}} = \sum_{r=1}^{R} W_r \circ H_r \circ Q_r. \tag{2}$$
- *Nonnegative Tucker Decomposition (NTD)* Unlike CP decomposition, which assumes a sum of rank-one components, Tucker decomposition introduces a core tensor $\mathcal{G}$ that interacts with factor matrices. For a third-order tensor, NTD is parametrized by four factors: $\mathcal{G} \in \mathbb{R}_+^{r_1 \times r_2 \times r_3}, W \in \mathbb{R}_+^{I \times r_1}, H \in \mathbb{R}_+^{J \times r_2}, Q \in \mathbb{R}_+^{K \times r_3}$ such as:
$$\mathcal{X} \approx \hat{\mathcal{X}} = \mathcal{G} \times_1 W \times_2 H \times_3 Q, \tag{3}$$
with dimension parameters r_1, r_2, r_3 inducing the low-rank aspect of the decomposition. This decomposition provides additional flexibility in modeling dependencies across different modes, because elements in factors may be combined. In that sense, it can be seen as an intuitive extension of dictionary learning and sparse coding to tensors.

  NTD has been employed to uncover structured and interpretable patterns in music [17]. This property may serve as a creative tool for music composition, especially in sampling, or as an interactive entertainment system. This potential has been realized through the development of a graphical interface that allows users to explore and manipulate these extracted patterns intuitively(**) [58].

  Both NCP and NTD models are computed solving the optimization problem $\arg\min_{\hat{\mathcal{X}} \geq 0} D(\mathcal{X} \| \hat{\mathcal{X}})$(***).

Deep nonnegative low-rank factorization models – Deep extensions of low-rank factorization are defined as a nonnegative low-rank factorization model where at least one mode is modeled by several nonnegative components, *e.g.*, deep NMF [30], where the first mode is modeled with n nonnegative matrices W_i, such that $X \approx W_1 W_2 ... W_n H$. To the best of our knowledge, deep NMF models have never been studied in the context of music processing, and deep alternatives to NPC and NTD have not yet been developed.

1.2.3 Deep Learning models

Definition – Deep learning is a subfield of machine learning that consists of algorithms applying a sequence of multiple nonlinear transformations, known as "layers" – the number of which gives rise to the term "deep". These algorithms are generally classified under the broader category of "Artificial Neural Networks", their historical name. Deep learning models, trained on large-scale datasets, have revolutionized MIR [20] due to

(**) https://unmixer.ongaaccel.jp/

(***) Here, the nonnegativity constraint $\hat{\mathcal{X}} \geq 0$ means that all factors parametrizing $\hat{\mathcal{X}}$ are nonnegative.

their ability to learn efficient representations directly from data, eliminating the need for handcrafted features or predefined structures.

Standard architectures – Deep learning models in MIR typically consist of:

1. *Pre-processing block:* Historically, raw audio was transformed into spectrograms (*e.g.*, STFT) in a deterministic operation. However, a few recent end-to-end models propose to integrate this step, using differentiable operations such as 1D convolutions to optimize feature extraction [48–51].

2. *Representation Learning block:* The core component, designed for learning efficient representations of music, denoted as "embeddings". This block is predominantly based on Convolutional Neural Networks (CNNs), Transformers, or combinations of both approaches.

   CNNs, introduced initially for computer vision [32], excel at learning spatial cues in images, which turn into rich and hierarchical features with depth. Despite some fundamental differences between spectrograms and natural images [20] – *e.g.*, the loss of translation invariance in the frequency domain –, such models proved particularly effective for spectrogram-based music analysis by interpreting spectrograms as images [2, 21, 46].

   Transformers, originally developed for natural language processing [59], are now gaining traction in MIR due to their ability to model long-range dependencies efficiently. Unlike CNNs (and the formerly used Recurrent Neural Networks), Transformers use self-attention mechanisms to weigh relationships between all time steps simultaneously, enabling them to capture complex, multi-scale musical structures [60].

3. *Output block:* The ouput block is task-dependent. For classification tasks (*e.g.*, genre recognition, instrument classification), the output layer often consists of a multi-layer perceptron [20]. For music source separation, the output is often a set of masks applied to the input spectrogram and/or to the waveform, reconstructing the separated audio signals [4, 21]. For music generation, the output can be a decoder network that reconstructs audio from learned embeddings [44, 45].

General-purpose representations – Recent deep learning models for MIR increasingly focus on learning general-purpose representations, *i.e.* embeddings that efficiently encode diverse musical attributes in order to be adaptable to multiple downstream tasks. Of particular interest in this proposal are foundation models [24, 25], which leverage self-supervised learning to train on large-scale, unlabeled music datasets. Inspired by self-supervised paradigms in speech and NLP, these models have demonstrated remarkable efficiency in learning task-agnostic embeddings, making them transferable across different MIR applications (*e.g.*, genre classification, key recognition, source separation [24]). A second example of models developed for general-purpose representations are neural audio codecs [44, 45], that aim to compress raw audio signals into compact embeddings while preserving perceptual fidelity. Nonetheless, because these neural audio codecs are developed for general audio and not only for music, their comprehension of the particularities of music may be lower than those of the aforementioned foundation models.

The key advantage of task-agnostic embeddings lies in their holistic understanding of music: rather than being tailored to a single task, general-purpose representations capture broad musical properties (*e.g.*, timbre, rhythm, pitch structure, harmonic progression) that can be leveraged for multiple applications. This reduces the reliance on large task-specific labeled datasets, a critical advantage given the limited availability of annotated musical data (notably due to copyrights).

Nonetheless, it may not always be possible to use directly such representations, for instance due to their large computation requirements, or because they may not help in addressing a specific problem. Hence, an intriguing line of work in deep learning models nowadays consists of post-processing a foundation model to compress it [39, 40] or adapt it, either for a new task [26, 27, 61] or for improving interpretability [62].

Hybrid deep learning models – An encouraging direction for future MIR research is to parametrize deep learning models with respect to signal processing knowledge and structures, coined as "model-based deep learning" [41]. In this proposal, we will use the terminology of "hybrid models" to mention model-based deep learning using principles of low-rank factorizations.

One first approach in hybrid models consists of incorporating low-rank factorizations directly into deep learning architectures, creating an explicit mathematical bridge between neural networks and factorization models [63, 64]. This structured design has the potential to promote interpretability and ease model control. However, while intuitive, directly enforcing low-rank constraints can be overly restrictive, potentially compromising model expressiveness. In particular, naive low-rank approximations may struggle to capture complex hierarchical structures, which are essential for deep learning models to generalize effectively. To address this limitation, recent work [42, 43] has explored more sophisticated architectures, demonstrating promising results in constructing structured yet expressive deep representations. These developments provide a compelling starting point for further investigation in this direction.

From a broader perspective, the idea of structured latent representations also appears in the domain of neural audio codecs [44, 45]. These models often employ vector quantization, a technique that discretizes latent representations, effectively enforcing a form of binary or quantized low-rank factorization. In this sense, neural audio codecs can be seen as a constrained form of hybrid models.

1.2.4 Tasks tackled in this proposal

While many audio tasks could be of interest, we will focus on a selected subset that has been extensively studied using low-rank factorization and previously explored by the PI, ensuring methodological continuity and facilitating evaluation.

Automatic Music Transcription (AMT) – AMT converts audio recordings into symbolic representations (e.g., MIDI), estimating discrete musical events such as pitch, onset, and duration. The complexity of AMT arises from the polyphonic nature of most music, where multiple sources produce overlapping overtones that blend in the frequency domain, making it difficult to distinguish individual notes [6]. In addition, because music sources are synchronized on the rhythm, most notes are played simultaneously. Furthermore, AMT must account for expressive variations such as vibrato, pitch bends, and dynamic changes. Finally, it is still very challenging to develop models that are agnostic of the instrument, with very recent attempts obtaining convincing performance [2]. On this latter point, because structured methods are less dependent on data, they are assumed to generalize better between instruments [6].

In musicology, AMT plays a crucial role in analyzing historical recordings, reconstructing lost musical scores, and studying stylistic evolution. For musicians, it enables automatic transcription of improvisations, aiding in learning and composition. AMT is fundamental in MIR, as it transforms raw audio data into a structured format suitable for further analysis, retrieval, and generation.

AMT was one of the first tasks where nonnegative low-rank factorizations were successively applied to music analysis [12, 18, 38]. Deep learning methods are nowadays the most efficient methods [2, 3, 6].

Music Source Separation (MSS) – MSS decomposes a mixed audio signal into individual sources (instruments). The difficulty in MSS arises from the fact that instruments often share overlapping frequency components and that the observed signal is a nonlinear sum of sound waves interacting in a complex acoustic environment (in particular due to the modifications due to the recording conditions and mixing) [7]. In addition, a challenge of MSS lies in accurately estimating source components while preserving perceptual quality, particularly in the presence of reverberation, background noise, and dynamic interplay between instruments.

MSS is essential in MIR for tasks such as remixing, instrument isolation, and spatial audio enhancement. From a musicological perspective, MSS may allow researchers to analyze individual performance elements in recordings – in particular for historical or live recordings, which may be difficult to understand as such – and facilitate in-depth studies of orchestration and timbre. Musicians may benefit from MSS when composing by sampling and generating accompaniment tracks, and may contribute to creating practice tools that isolate specific instruments for targeted learning.

Low-rank factorization methods are an important tool for source separation, with many models devoted to it. A detailed overview may be found in [7, Chap. 8, 9, 16]. Nonetheless, the most recent developments are generally deep learning models [4, 21], with unmatched performance.

Music Structure Analysis (MSA) – MSA consists of estimating the overall organization of a song, *i.e.* segmenting a musical piece into sections – such as verses, choruses, and bridges – while also identifying recurring motifs and thematic elements – the content of these sections. A difficulty in music structure analysis is the fact that structure is ambiguous and hard to define, as it can be influenced by many musical characteristics: harmonic progressions, melodic patterns, timbral changes, and rhythmic cues, among others [11]. An additional complexity of MSA arises from the hierarchical nature of music structure, where small-scale repetitions (*e.g.*, motifs, patterns) coexist with large-scale structures (*e.g.*, sonata form, fugue, 12-bar blues).

MSA is particularly useful in musicology, where it enables the systematic analysis of compositional styles and the comparison of musical pieces based on structure rather than surface-level characteristics. For musicians, MSA may aid in practice and performance planning by highlighting key transitions and repeated sections. In MIR, structure analysis may also be seen as an intermediate task, and useful for applications such as music summarization, cover song detection, and automated playlist generation.

MSA has been studied using both nonnegative low-rank factorization [16, 17, 64] and deep learning [22, 23] methods, with competitive performance from both methods on the evaluated datasets.

1.3 Methodology and Risk Management

We first introduce the work packages in Sections 1.3.1 to 1.3.4, derived from research questions introduced in Section 1.1.2, as shown in the table below. Then, we present ethical considerations in Section 1.3.5. Finally, Sections 1.3.6 to 1.3.8 present the deliverables, risk management, and temporal organization of the project.

	Obj. 1 & 2			Obj. 3
	RH1	RH2	RH3	
WP1	×			~
WP2		×		~
WP3			×	×
WP4				×

1.3.1 WP1 – Integrating Deep Learning Principles to Nonnegative Low-Rank Factorization

This work package aims to integrate principles from deep learning into the design of structured models, with the objective of enhancing the expressivity and scalability of nonnegative low-rank factorization methods while preserving their inherent interpretability and steerability. In practice, we will develop and evaluate deep versions of NMF, NTD, and NCP and explore how low-rank models can be adapted for large-scale learning by leveraging batching, epochs, and other training procedures inspired by modern deep learning pipelines (including GPU-based computation, as supported by toolboxes such as TensorLy [65]). The anticipated benefits of depth and large-scale learning are outlined in Research Hypothesis 1, presented in Section 1.1.2.

Preliminary work

The PI is currently developing an initial benchmark for low-rank factorization methods [66], which is expected to integrate the algorithms from [30] in the near future. In this context, the PI is collaborating with Dr. Leplat to implement a Python version of the algorithms introduced in [30] for integration into the benchmark.

Tasks

- WP1.1 – Evaluate the performance of deep NMF models in MIR

  As a first task, we focus on implementing and benchmarking the deep NMF models introduced in [30]. We will conduct extensive evaluations comparing deep and shallow NMF variants and deep learning baselines in AMT, MSS, and MSA. Special attention will be paid to the interpretability of the learned representations across layers, assessing whether deep NMF introduces meaningful hierarchies in time-frequency content. Additionally, we will investigate how constraints such as sparsity, harmonicity [7], and minimum-volume [30] can enhance the representations obtained by the model.

- WP1.2 – Develop algorithms for deep nonnegative tensor factorization methods

  In this task, we aim to extend deep factorizations to tensor-based models, focusing on Nonnegative Tucker Decomposition (NTD) and Nonnegative Canonical Polyadic (NCP) decomposition. We will design novel deep versions of NTD and NCP by stacking multiple factor layers on one or several tensor modes. To this end, we will adapt existing optimization strategies, starting with an approach based on matricizing the tensor product – as proposed in [17] – which allows us to leverage algorithms initially developed for deep NMF [30]. As this strategy may prove computationally demanding, we may consider alternative formulations in a second stage, aiming to reduce computational complexity. These models will be applied to higher-dimensional representations of music signals, such as time-frequency–channel [14] or time-frequency–bar [17] tensors, enabling richer modeling of spectral, spatial, and/or structural patterns. Particular attention will be paid to the trade-offs between expressivity, interpretability, and computational cost, especially given that even shallow NTD and NCP models already carry significant computational overhead.

- WP1.3 – Investigate scaling (shallow and deep) nonnegative low-rank models by leveraging large datasets

  A third task, complementary to both previous ones, explores how nonnegative low-rank models – both shallow and deep – can be adapted for large-scale learning, drawing on training techniques used in deep learning. We will introduce batch-wise and epoch-based training procedures into the optimization of factorization models, a strategy infrequently used in traditional structured methods (though not entirely new [37]) – large-scale learning generally focuses on accelerating and/or parallelizing algorithms to process more data at-once [67, 68]. We will assess how these modifications affect model performance and interpretability and compare large-scale variants of NMF, NTD, and NCP to deep learning baselines on the same data and tasks.

1.3.2 WP2 – Interpretation-Driven Model Adaptation and Compression

This work package investigates deep learning models for MIR, in particular foundation models [24, 25], focusing on interpretability and steerability. Our approach centers on model adaptation and model compression, two rapidly advancing research areas that have demonstrated remarkable success in their respective objectives – adapting models to new settings and reducing model complexity while maintaining performance. We will explore how these techniques can be repurposed to extract structured, interpretable representations from otherwise opaque models.

Tasks

- WP2.1 – Improving interpretability in pre-trained models via adaptation techniques

  In this task, we will investigate whether parameter-efficient adaptation techniques can enhance the interpretability of pre-trained deep learning models for music, such as foundation models [24, 25]. We will focus in particular on low-rank adaptation methods like LoRA [26] and their tensor-based extensions [61], which introduce structured, low-rank updates without modifying the full model. By acting as inductive biases, these techniques may guide learned representations toward musically meaningful features – an outcome that has already been demonstrated in previous work on model adaptability [62]. Where applicable, we will also test whether adaptation can improve steerability by allowing explicit control over model behavior in inference.

- WP2.2 – Enhancing interpretability through compression techniques

  This task examines whether compressing deep models using nonnegative low-rank factorization can also promote interpretability, beyond simply reducing the model size. The idea stems from the nature of nonnegative low-rank models themselves: as dimensionality reduction techniques, they emphasize the most salient structures in data. In music, their additive structure tends to uncover interpretable frequency templates. While most existing work has focused on the performance gains of low-rank compression [39, 40], we hypothesize that introducing structured constraints – nonnegativity, but also sparsity and harmonicity – into deep learning models can encourage internal representations that are more interpretable.

AAPG2025	MusAIc		JCJC
Coordinated by:	Axel MARMORET	48 months	312,125€
Theme E.2 – Artificial intelligence and data science			

1.3.3 WP3 – Hybrid Deep Learning Models: Steerable, Interpretable, and Efficient Architectures by Design

This work package aims to develop hybrid models that are interpretable and steerable by design. These models will combine the structured, interpretable representations of nonnegative low-rank factorization methods with the expressive power of deep learning. While this work package will build on insights from WP1 and WP2, it will also begin with the exploration of neural audio codecs [44, 45] – already promising for efficient and structured representations – as an initial foundation, even in the absence of prior results. By embedding structured constraints directly into the model architecture, WP3 seeks to develop models that are high-performing, steerable and interpretable. This work package is the most promising in terms of translating the project's outcomes into usable tools for music professionals, in line with Objective 3.

Tasks

- WP3.1 – Adapt neural audio codecs for interpretability purposes

  This task investigates how architectures inspired by neural audio codecs [44,45] can serve not only compression purposes but also interpretability and steerability goals. Neural audio codecs already encode audio signals into compact latent representations using quantization or structured bottlenecks. We will explore how injecting structured priors—such as nonnegativity, sparsity, or harmonicity—during training might encourage semantic decompositions in the latent space, and whether vector quantization can align with musically meaningful elements such as note templates, timbre, rhythm, or motifs. We will also assess the extent to which interpretable control can be exercised over audio generation or reconstruction, for example, by manipulating the latent space directly, as in DDSP [50]. This task will serve as a first concrete step toward bridging structured and data-driven models, even before incorporating insights from WP1 and WP2.

- WP3.2 – Leverage knowledge from WP1-2 to develop controllable by-design hybrid models

  This task aims to develop hybrid deep learning models that are interpretable and steerable by design, building directly on the insights gained in WP1 and WP2. Rather than retrofitting interpretability or adding isolated constraints to existing architectures, we will explore how to embed these properties natively within model design. At this stage, we intentionally leave the architectural choices open in order to remain receptive to novel approaches that may emerge over the course of the project. Nonetheless, recent advances [42, 43] demonstrate promising strategies for integrating low-rank constraints and structured representations into deep networks. This task also aligns with the broader model-based deep learning paradigm advocated by Richard *et al.* [41], and directly supports the overarching objective of the MusAIc project to produce AI models that are high-performing, steerable, and interpretable.

  WP3.2 operationalizes Research Hypothesis 3 by synthesizing the outcomes of WP1 and WP2 into unified hybrid architectures. It also directly contributes to Objective 3 by moving toward models that are both controllable and usable by music professionals.

1.3.4 WP4 — Dissemination, and Co-Design with Music Professionals

To ensure real-world impact, AI models must be usable and valuable for the communities they aim to serve. WP4 focuses on transferring the project's outcomes to academic circles and beyond by involving musicians, musicologists, and sound engineers in the co-design and evaluation of the developed tools. In addition to contributions to the academic domain through open-source publications and code, we plan to engage with a consortium of music professionals and students, including those already identified in Section 2.1.2, as well as new collaborators met through conferences, presentations, and outreach efforts.

Tasks

- WP4.1 — Iterative Co-Design with Music Professionals *(Sparse contributions for* $24m$*)*

  The first task of this work package is to establish an ongoing, collaborative workflow with the music consortium to embed feedback from professionals throughout the development process. Through regular meetings (every month or two), interviews, informal discussions, and practical tests, we aim to better understand how notions of

interpretability and steerability are perceived and defined by different musical communities. These interactions will help identify practical use cases, reveal potential usability gaps in the models, and guide the refinement of design choices. The co-design process is expected to begin in the second half of the project, once initial models are sufficiently mature, and will continue as a form of iterative validation.

- WP4.2 — Open Dissemination and Knowledge Transfer

  The second task centers on ensuring that the project's scientific and technological outcomes are broadly accessible, reusable, and impactful. This includes the publication of research articles in top-tier conferences and journals (*e.g.*, ISMIR, ICASSP, ICML, NeurIPS, TISMIR, JMLR), the release of open-source software and benchmark datasets, and the contribution to standardization efforts within the MIR and ML communities. In line with our commitment to open science, we will also produce educational resources and documentation – such as tutorials, web-based demonstrations, or reproducible notebooks – that help make our models understandable and usable by a broader audience. This task will run continuously throughout the project and serve as a key interface between the research, educational, and creative communities.

1.3.5 Ethical considerations

Ecological considerations

Training large-scale deep learning models from scratch has a significant carbon footprint. To mitigate this impact, we will rely on pre-trained models in WP2 and aim to reduce model sizes in WP1 and WP3. Our approach seeks a balance between increasing depth for mathematical expressivity and limiting computational costs for ecological reasons.

Concerns on open science

Open science is a foundational principle of this project. All publications produced during the project will be uploaded to HAL, and we will prioritize open-access venues such as ISMIR, ICML, NeurIPS, TISMIR, and JMLR. Our code will be released under open-source licenses, following the PI's precedent in previous work [66, 69], and we plan to contribute to existing toolboxes, such as TensorLy [65], particularly for low-rank factorization. We also plan to distribute our models through the web portal of the research program "*Analyse musicale assistée par ordinateur*", which aims to catalog existing tools for music analysis.

While our primary focus is music analysis, the developed models may apply to other audio domains (*e.g.*, ecoacoustics, speech) and broader fields where low-rank factorization has proven valuable (*e.g.*, hyperspectral imaging, neuroimaging). In this context, long-term accessibility and maintainability of models and tools are essential.

When it comes to datasets, we will favor open data whenever possible. Where copyright constraints prevent this (common in music), we will use widely accepted standard datasets to ensure comparability and reproducibility. To support benchmarking, the PI has led the development of a reproducible evaluation framework for low-rank factorization on audio signals [66], bridging the audio and low-rank communities.

This open, transparent, and community-oriented approach fully aligns with ANR's priorities in open science.

Alignment with the BRAIn Team Philosophy

These commitments reflect the philosophy of the BRAIn team (Better Representations for Artificial Intelligence)[(****)], to which the project coordinator belongs. The team promotes the development of AI models that are not only high-performing but also resource- and data-efficient, accessible, and reproducible. Since its creation, the BRAIn team has actively supported open publishing practices: all articles are uploaded to HAL, source code is systematically released on the team's GitHub[(#)], and datasets are shared under open licenses (e.g., the Silent Cities dataset [70] is available under a CC-BY license).

1.3.6 Deliverables

- $T_0 + 6m$ — Data Management Plan.

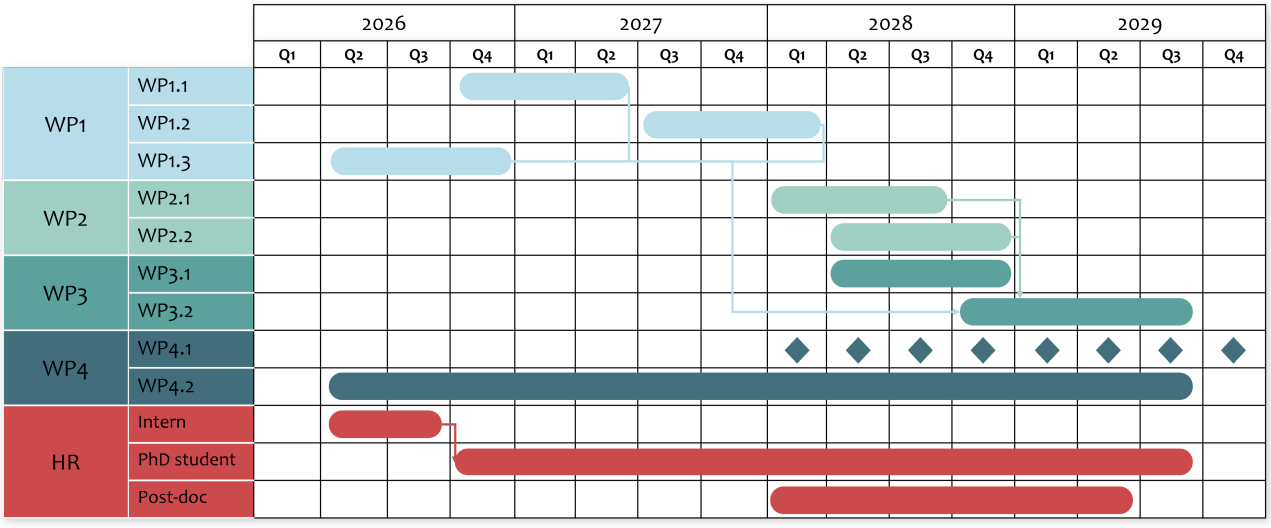
[(****)] https://www.imt-atlantique.fr/en/research-innovation/teams/brain
[(#)] https://github.com/brain-bzh/

- $T_0 + 30m$ — Intermediate report on findings with relation to WP1. This includes communication of the code to music professionals if it has not already been shared earlier in the project.
- $T_0 + 39m$ — Intermediate report on findings with relation to WP2. This includes communication of the code to music professionals if it has not already been shared earlier in the project.
- $T_0 + 48m$ — Final report, summarizing impact of previous work, plus new findings with relation to WP3. This includes communication of the code to music professionals if it has not already been shared earlier in the project.
- *Continuous* — Presentation of the work in public events (*e.g.*, Ressac festival[(##)], Le Mans Sonore[(###)], Fête de la Science [(####)], and other opportunities).

1.3.7 Risks & Fallback solutions

- Late recruitment of the PhD student — We plan to recruit an intern who may transition into a PhD position. However, if the intern proves not to be a good fit, the PhD recruitment process could become more challenging. In the event of a delayed PhD recruitment, permanent researchers will handle WP1.1, allowing the newly hired PhD student to begin directly with WP1.2. To mitigate this risk, we have already started outreach to targeted programs, including the Computer Science curriculum at IMT Atlantique (where the PI teaches), the ATIAM master's program at Sorbonne University, and international music technology programs (*e.g.*, NYU, Universitat Pompeu Fabra, Queen Mary University).
- Late recruitment of the post-doc researcher — In that situation, WP2 will be delayed, and the first contributions to WP3 would be based on WP1. In case of very late recruitment, the PhD student may also be interested in pursuing the work after their defense. To avoid this situation, we will communicate the position through our research networks.
- Impossibility to address WP3.2 — Most of the WPs in this project are independent, with the exception of WP3.2 that depend on outcomes of WP1 and WP2. If the findings of these tasks are not sufficient to address WP3.2, the PhD student and the post-doc researcher will focus on WP3.1 and tackle WP4 using existing tools.

1.3.8 Temporal organization of the project



2 Organisation and implementation of the project

(##) https://www.univ-brest.fr/festival-ressac/
(###) https://lemanssonore.fr/
(####) https://www.fetedelascience.fr/

AAPG2025	MusAIc		JCJC
Coordinated by:	Axel MARMORET	48 months	312,125€
Theme E.2 – Artificial intelligence and data science			

2.1 SCIENTIFIC COORDINATOR AND ITS TEAM

2.1.1 Scientific coordinator

Axel MARMORET

After completing my Ph.D. in 2022 at IRISA (Université de Rennes 1), I joined in 2023 the BRAIn team at IMT Atlantique (Lab-STICC) as an Associate Professor. My research lies at the intersection of AI and audio signal processing, with a particular focus on MIR. I have contributed both methodological advances – proposing novel paradigms for music analysis [17, 54, 71] – and algorithmic developments, notably through open-source toolboxes [66, 69]. Music has always been central in my life – I have been a musician for nearly 20 years – and this personal connection deeply shapes how I approach research. While music remains my main focus, I have also begun exploring other sound-related domains, including ecoacoustics [72].

Since my recruitment, I have been supervising three ongoing PhD theses, including one funded by the ANR ENDIVE project led by B. Pasdeloup, a colleague in the BRAIn team. As PI of the MusAIc project, I will be responsible for identifying key research questions and defining experimental protocols to address them.

2.1.2 Coordinator's team

In this project, I will be supported by a carefully assembled team of experts whose complementary perspectives cover all key dimensions of the work. Each brings their own specific interests and expertise, enriching the scientific dialogue throughout the project. The team's strengths span both the technical aspects of AI and signal processing, as well as the scientific and artistic dimensions of music.

Nicolas FARRUGIA

Nicolas is a Professor at IMT Atlantique, Lab-STICC, in the BRAIn team. He is a researcher in audio and neurocognition, with a strong deep-learning background. He is also a skilled musician. He will bring important insight to the project when discussing deep learning for audio signals processing [73, 74] as well as an artistic vision. Nicolas will direct the PhD position.

Vincent GRIPON

Vincent is a Professor at IMT Atlantique, Lab-STICC, and the head of the BRAIn team. He conducts research in many aspects of deep learning, in particular model compression and adaptation [27, 75], and will provide an expert eye in WP2.

Nicolas GILLIS

Nicolas is a Full Professor at the University of Mons in Belgium. He obtained his PhD from the Université catholique de Louvain (UCL) in 2011, for which he was awarded the Householder Prize in 2014. He received an ERC Starting Grant in 2015 and an ERC Consolidator Grant in 2023 to work on constrained low-rank matrix approximations and their extensions.

His research focuses on numerical linear algebra, optimization, signal processing, and machine learning. A central topic in his research is NMF, on which he published a book in 2020 [55]. He recently contributed to the development of deep NMF [30]. Thus, Nicolas will play a key role in the development of WP1, notably because this work package builds on and complements the themes explored in Nicolas' ERC grants.

Gaël RICHARD

Gaël is a Full Professor in Télécom Paris in the field of audio signal processing. He is also the co-scientific director of the Hi! PARIS interdisciplinary center on AI and Data analytics. In 2020, he received the Grand Prize of IMT-National Academy of Science. In 2022, he was awarded an advanced ERC grant from the European Union for a project on hybrid artificial models for sound.

His research interests are mainly in the field of speech and audio signal processing and include topics such as source separation, machine learning methods for audio/music signals, and MIR. His most recent works are dedicated to interpretable AI models [42, 62], along with an emphasis on developing hybrid models [41]. Thus, Gaël will play a key role in the context of WP2 and WP3.

Multidisciplinary Consortium – artistic partners

Some music professionals, listed hereafter, have already accepted the invitation to participate in the music consortium. We plan on involving new partners throughout this project, encountered through conferences, presentations, and public outreach efforts.

- Christophe GUILLOTEL-NOTHMANN is a musicology researcher at the CNRS, whithin the *Institut de Recherche en Musicologie*. Among his various research directions, some explore the interface between musicology and computer science – such as the program "*Analyse musicale par ordinateur*", which focuses on cataloguing and developing software for music analysis, and the collaborative annotation interface Tonalities, designed to support the use and enrichment of open music datasets.
- Arthur LAUTH is a sound engineer and musician. Arthur has had his own recording studio ("Brown Bear Recordings") since 2009, and he is *intermittent du spectacle* since 2012.
- Adrien LLAVE is a researcher at Orange on topics related to AI and audio signal processing. He is also a trained sound engineer, having graduated from the *ENS Louis Lumière* in 2015. Finally, Adrien is a musician.

We also plan to collaborate closely with IRCAM, whose mission aligns with this project's objectives. In addition, we have initiated contact with the Brest Conservatoire and hope to involve them actively in the project.

2.1.3 Hired personnel

Master's/PhD student

The master's student will begin by addressing WP1.3, which presents relatively low risk and offers a solid entry point into the project. Upon continuation as a PhD student, they are expected to focus initially on WP1.1, before progressively contributing to WP1.2, WP3.1, and finally, WP3.2. Throughout both the master's and doctoral phases, the student will also be actively involved in the continuous integration and dissemination efforts outlined in WP4.

A good profile for a candidate would be a master's student or engineer with a strong mathematical and signal processing background. The anticipated ideal master's/PhD student would have studied in the ATIAM master (Sorbonne University) or similar music technology masters in worldwide universities (*e.g.*, NYU, Universitat Pompeu Fabra, Queen Mary University).

Post-doc

The post-doctoral researcher will primarily contribute to WP2, with subsequent involvement in WP3.2. They will also play an active role in the dissemination and collaborative activities outlined in WP4.

The ideal candidate for this position is a researcher with prior experience in advanced deep learning models, particularly within the context of MIR. Starting the work at the beginning of 2028 has been chosen to correspond to standard PhD defense dates at the end of 2027, which also allows them to align with advances obtained by the PhD student so that they contribute together to WP3.2.

2.2 Implemented and requested resources to reach the objectives

Staff expenses

As detailed in Section 2.1.3, I plan to hire a master's student (6 months, 4.5k€) expected to follow as a PhD student (36 months, 148k€); and a post-doc researcher (18 months, 90k€).

Instruments and material costs

The post-doc and PhD student will both receive a laptop and a screen for daily work (3k€ each).

Outsourcing costs

We require 5k€ publication fees for open-access publication.

Overhead costs

We require 11.5k€ to cover attendance to conferences at which articles will be published (anticipating 3 international conferences and 2 national conferences). In addition, we require 4k€ to organize research stays with the research partners at the University of Mons and Télécom Paris for the PhD student and the post-doc

respectively, but also for the PI. Visits should occur at key points of the project, *i.e.* at the beginning to start efficiently, and close to publication deadlines. Finally, we require 6k€ to organize meetings with the music consortium, in particular to account for travel costs. These meetings will be determinant in receiving feedback and organizing tests with music professionals.

Requested means by item of expenditure and by partner

		Partner 1 – IMT Atlantique
Staff expenses, including costs of a partial release from teaching obligations	Permanent (62 p.m.)	242,363€
	Non-permanent (60 p.m.)	242,500€
Instruments and material costs		6,000€
Building and ground costs		0€
Outsourcing / subcontracting		5,000€
Overhead costs	Travel costs	21,500€
	Administrative management & & structure costs	37,125€
Sub-total		**575,472.97€**
Requested funding		**312,125€**

3 Impact and benefits of the project

Scientific dissemination

This project addresses a significantly underexplored dimension of recent AI models in MIR: model interpretability and steerability. By focusing on these aspects, it aims to raise awareness of the importance of model transparency and controllability, and open new research directions within the broader machine learning community. Given its interdisciplinary nature, the project is expected to contribute to multiple scientific domains – including machine learning and MIR, but also musicology and performing arts – and engage a wide range of researchers.

Practically, the project will contribute to open-source libraries - particularly in connection with WP4.2 (Section 1.3.4) – to facilitate the adoption of developed methods by other researchers. In line with the open science policy of the BRAIn team, all code and experiments associated with project publications will be released on GitHub, ensuring reproducibility and fostering collaborative development.

Societal impact

Beyond its scientific contributions, the MusAIc project has a strong potential for societal impact, particularly through its emphasis on making advanced AI tools accessible, interpretable, and steerable to a broad audience of music professionals, students, and enthusiasts. By explicitly designing models that are steerable and interpretable, the project intends to respond to the practical needs of musicians, musicologists, sound engineers, and educators – groups who often face barriers when engaging with opaque or rigid AI systems.

In practical terms, this transfer will be facilitated through dedicated collaborations with a consortium of artistic partners and music institutions and the design of user-facing tools that integrate feedback from professionals. Ultimately, MusAIc strives to democratize access to high-performance AI technologies in music, empowering creative practices and enabling new forms of artistic exploration.

Impact on the PI position in the laboratory

This project represents a decisive step in shaping my long-term research identity, both within my laboratory and in the broader research community. At Lab-STICC, no current member specializes in interpretable and steerable AI models. By filling this gap, I will establish myself as the local reference on these critical issues, fostering new collaborations across research axes and positioning myself as a key contributor to cross-disciplinary initiatives. Within the BRAIn team, my work will complement existing expertise by addressing underexplored dimensions of deep learning, particularly the development of models designed for transparency, steerability, and public engagement. This aligns directly with the team's mission of building better, more responsible AI representations.

This project also offers a unique context to collaborate with leading experts such as Nicolas Gillis and Gaël Richard, whose respective contributions to low-rank factorization and MIR are internationally recognized. More broadly, it represents a decisive step in the development of my research agenda. Beyond advancing music analysis, I aim to explore how AI can support artistic expression in a wider sense – for tasks such as music generation, production, and creative interaction. If successful, MusAIc would not only launch my independent research career, but also lay the foundation for more ambitious projects in the future (*e.g.*, an ERC grant).

More broadly, I see MusAIc as the starting point of a long-term research program aimed at shaping the future of the MIR field. Situated at the crossroads of signal processing, AI, and the arts, MIR is uniquely positioned to reflect on the role of AI in creative practices. At a time when AI is often perceived as a threat by artistic communities, MusAIc's focus on interpretability and steerability gives it both scientific and ethical significance.

4 REFERENCES RELATED TO THE PROJECT

[1] J. S. Downie. Music information retrieval. *Annual review of information science and technology*, 37(1):295–340, 2003.
[2] R. M. Bittner, J. J. Bosch, D. Rubinstein, G. Meseguer-Brocal, and S. Ewert. A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. In *ICASSP 2022-2022 IEEE Int. Conf. Acoustics, Speech, Signal Process.*, pages 781–785. IEEE, 2022.
[3] R. Wu, X. Wang, Y. Li, W. Xu, and W. Cheng. Piano transcription with harmonic attention. In *ICASSP 2024-2024 IEEE Int. Conf. Acoustics, Speech, Signal Process.*, pages 1256–1260. IEEE, 2024.
[4] S. Rouard, F. Massa, and A. Défossez. Hybrid transformers for music source separation. In *ICASSP 2023-2023 IEEE Int. Conf. Acoustics, Speech, Signal Process.* IEEE, 2023.
[5] Z. Evans et al. Stable audio open. In *ICASSP 2025-2025 IEEE Int. Conf. Acoustics, Speech, Signal Process.* IEEE, 2025.
[6] E. Benetos, S. Dixon, Z. Duan, and S. Ewert. Automatic music transcription: An overview. *IEEE Signal Process. Mag.*, 36(1):20–30, 2018.
[7] E. Vincent, T. Virtanen, and S. Gannot. *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.
[8] M. McVicar, R. Santos-Rodríguez, Y. Ni, and T. De Bie. Automatic chord estimation from audio: A review of the state of the art. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 22(2):556–575, 2014.
[9] J. Salamon, E. Gómez, D. P. Ellis, and G. Richard. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Process. Mag.*, 31(2):118–134, 2014.
[10] M. F. Matthew E. P. Davies, Sebastian Bock. *Tempo, Beat and Downbeat Estimation*. https://tempobeatdownbeat.github.io/tutorial/intro.html, Nov. 2021.
[11] O. Nieto et al. Audio-based music structure analysis: Current trends, open challenges, and applications. *Trans. Int. Society Music Information Retrieval (TISMIR)*, 3(1), 2020.
[12] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *2003 IEEE Workshop Applications Signal Process. Audio Acoustics (WASPAA)*, pages 177–180. IEEE, 2003.
[13] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, 21(3):793–830, 2009.
[14] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Trans. Audio, Speech, Language Process.*, 18(3):550–563, 2009.
[15] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama. Complex NMF: A new sparse representation for acoustic signals. In *ICASSP 2009-2009 IEEE Int. Conf. Acoustics, Speech, Signal Process.*, pages 3437–3440. IEEE, 2009.
[16] O. Nieto and T. Jehan. Convex non-negative matrix factorization for automatic music structure identification. In *ICASSP 2013-2013 IEEE Int. Conf. Acoustics, Speech, Signal Process.*, pages 236–240. IEEE, 2013.
[17] A. Marmoret, J. Cohen, N. Bertin, and F. Bimbot. Uncovering audio patterns in music with nonnegative Tucker decomposition for structural segmentation. In *ISMIR*, pages 788–794, 2020.
[18] H. Wu, A. Marmoret, and J. E. Cohen. Semi-supervised convolutive nmf for automatic music transcription. In *Proc. 19th Sound and Music Computing Conf.*, 2022.
[19] M. Müller. *Fundamentals of music processing: Audio, analysis, algorithms, applications*, volume 5. Springer, 2015.
[20] G. Peeters and G. Richard. Deep learning for audio and music. In *Multi-faceted Deep Learning: Models and Data*, pages 231–266. Springer, 2021.
[21] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal Open Source Software*, 5(50):2154, 2020.
[22] T. Grill and J. Schlüter. Music boundary detection using neural networks on combined features and two-level annotations. In *ISMIR*, pages 531–537, 2015.
[23] M. Buisson, B. McFee, S. Essid, and H. C. Crayencour. Self-supervised learning of multi-level audio representations for music segmentation. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2024.
[24] Y. Li et al. Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*, 2023.
[25] M. Won, Y.-N. Hung, and D. Le. A foundation model for music informatics. In *ICASSP 2024-2024 IEEE Int. Conf. Acoustics, Speech, Signal Process.*, pages 1226–1230. IEEE, 2024.
[26] E. J. Hu et al. LoRA: Low-rank adaptation of large language models. In *Int. Conf. Learning Representations (ICLR)*, 2022.
[27] Y. Bendou, A. Ouasfi, V. Gripon, and A. Boukhayma. Proker: A kernel perspective on few-shot adaptation of large vision-language models. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2025.
[28] G. Wilson and D. J. Cook. A survey of unsupervised deep domain adaptation. *ACM Trans. Intelligent Systems Technology*, 2020.
[29] R. Dwivedi et al. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9), 2023.
[30] V. Leplat, L. T. K. Hien, A. Onwunta, and N. Gillis. Deep nonnegative matrix factorization with beta divergences. *Neural Computation*, 36(11):2365–2402, 2024.
[31] Y. Levine, N. Wies, O. Sharir, H. Bata, and A. Shashua. Limits to depth efficiencies of self-attention. *Advances in Neural Information Processing Systems*, 33:22640–22651, 2020.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
[33] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
[34] R. Eldan and O. Shamir. The power of depth for feedforward neural networks. In *Conf. Learning Theory*. PMLR, 2016.
[35] N. Cohen, O. Sharir, and A. Shashua. On the expressive power of deep learning: A tensor analysis. In *Conf. Learning Theory*, pages 698–728. PMLR, 2016.
[36] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conf. Computer Vision*, pages 818–833. Springer, 2014.
[37] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Int. Conf. Machine Learning (ICML)*, pages 689–696, 2009.
[38] E. Vincent, N. Bertin, and R. Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. Audio, Speech, Language Process.*, 18(3):528–537, 2009.
[39] A.-H. Phan et al. Stable low-rank tensor decomposition for compression of convolutional neural network. In *European Conf. Computer Vision*, pages 522–539. Springer, 2020.
[40] Y. Zniyed, T. P. Nguyen, et al. Hybrid network compression through tensor decompositions and pruning. In *32nd European Signal Process. Conf.*, 2024.
[41] G. Richard, V. Lostanlen, Y.-H. Yang, and M. Müller. Model-based deep learning for music information research: Leveraging diverse knowledge sources to enhance explainability, controllability, and resource efficiency. *IEEE Signal Process. Mag.*, 2025.
[42] J. Parekh, S. Parekh, P. Mozharovskyi, G. Richard, and F. d'Alché Buc. Tackling interpretability in audio classification networks with non-negative matrix factorization. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2024.
[43] M. Lebourdais, T. Mariotte, A. Almudévar, M. Tahon, and A. Ortega. Explainable by-design audio segmentation through non-negative matrix factorization and probing. In *Interspeech 2024*, 2024.
[44] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.
[45] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 30:495–507, 2021.
[46] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello. Deep salience representations for f0 estimation in polyphonic music. In *ISMIR*, pages 63–70, 2017.
[47] N. Sturmel et al. Linear mixing models for active listening of music productions in realistic studio conditions. In *Audio Engineering Society Convention 132*, 2012.
[48] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.
[49] M. Ravanelli and Y. Bengio. Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop (SLT)*, pages 1021–1028. IEEE, 2018.
[50] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts. DDSP: Differentiable Digital Signal Processing. In *Int. Conf. Learning Representations (ICLR)*, 2020.
[51] G. Peeters, G. Meseguer-Brocal, A. Riou, and S. Lattner. *Deep Learning 101 for Audio-based MIR, ISMIR 2024 Tutorial*. 2024.
[52] D. Haider, V. Lostanlen, M. Ehler, and P. Balazs. Instabilities in convnets for raw audio. *IEEE Signal Process. Lett.*, 2024.
[53] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
[54] A. Marmoret, F. Voorwinden, V. Leplat, J. E. Cohen, and F. Bimbot. Nonnegative tucker decomposition with beta-divergence for music structure analysis of audio signals. In *GRETSI 2022: XXVIIIe Colloque*. GRETSI, 2022.
[55] N. Gillis. *Nonnegative matrix factorization*. SIAM, 2020.
[56] N. Bertin. *Les factorisations en matrices non-négatives. Approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique.* PhD thesis, Télécom ParisTech, 2009.
[57] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.
[58] J. B. L. Smith, Y. Kawasaki, and M. Goto. Unmixer: An interface for extracting and remixing loops. In *ISMIR*, 2019.
[59] A. Vaswani et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
[60] C.-Z. A. Huang et al. Music transformer: Generating music with long-term structure. *arXiv preprint arXiv:1809.04281*, 2018.
[61] S. Jie and Z.-H. Deng. Fact: Factor-tuning for lightweight adaptation on vision transformer. In *Proc. AAAI Conf. Artificial Intelligence*, volume 37, pages 1060–1068, 2023.
[62] J. Parekh, S. Parekh, P. Mozharovskyi, F. d'Alché Buc, and G. Richard. Listen to interpret: Post-hoc interpretability for audio networks with nmf. *Advances in Neural Information Processing Systems*, 35:35270–35283, 2022.
[63] P. Smaragdis and S. Venkataramani. A neural network alternative to non-negative audio models. In *ICASSP 2017-2017 IEEE Int. Conf. Acoustics, Speech, Signal Process.*, pages 86–90. IEEE, 2017.
[64] A. Marmoret. *Unsupervised Machine Learning Paradigms for the Representation of Music Similarity and Structure.* PhD thesis, Université Rennes 1, 2022.
[65] J. Kossaifi, Y. Panagakis, A. Anandkumar, and M. Pantic. Tensorly: Tensor learning in python. *Journal Machine Learning Research (JMLR)*, 20(26), 2019.
[66] A. Marmoret. nmf_audio_benchmark, 2024.
[67] C. Liu, H.-c. Yang, J. Fan, L.-W. He, and Y.-M. Wang. Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce. In *Proc. 19th Int. Conf. World Wide Web*, pages 681–690, 2010.
[68] G. Chennupati, R. Vangara, E. Skau, H. Djidjev, and B. Alexandrov. Distributed non-negative matrix factorization with determination of the number of latent features. *The Journal of Supercomputing*, 76:7458–7488, 2020.
[69] A. Marmoret and J. E. Cohen. nn_fac: Nonnegative Factorization techniques toolbox, 2020.
[70] S. Challéat, N. Farrugia, J. S. Froidevaux, A. Gasc, and N. Pajusco. A dataset of acoustic measurements from soundscapes collected worldwide during the covid-19 pandemic. *Scientific Data*, 11(1):928, 2024.
[71] A. Marmoret, J. E. Cohen, and F. Bimbot. Barwise music structure analysis with the correlation block-matching segmentation algorithm. *Trans. Int. Society Music Information Retrieval (TISMIR)*, 6(1), 2023.
[72] A. Marmoret, N. Farrugia, and D. Cazau. Naïve unsupervised bioacoustics signal processing with nonnegative matrix factorization. In *2èmes Journées des Jeunes Bioacousticien·nes*, 2024.
[73] N. Farrugia, K. Jakubowski, R. Cusack, and L. Stewart. Tunes stuck in your brain: The frequency and affective evaluation of involuntary musical imagery correlate with cortical structure. *Consciousness and cognition*, 35:66–77, 2015.
[74] I. Moummad, R. Serizel, and N. Farrugia. Pretraining representations for bioacoustic few-shot detection using supervised contrastive learning. In *DCASE 2023-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2023.
[75] Y. Hu, S. Pateux, and V. Gripon. Adaptive dimension reduction and variational inference for transductive few-shot classification. In *Int. Conf. Artificial Intelligence Statistics*, pages 5899–5917. PMLR, 2023.