

Rossetto, Federico (2026) *Text augmentations from music artist biographies improve performance on genre classification, emotion recognition and music tagging.* PhD thesis.

https://theses.gla.ac.uk/86297/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses
https://theses.gla.ac.uk/
research-enlighten@glasgow.ac.uk

Text augmentations from music artist biographies improve performance on genre classification, emotion recognition and music tagging

Federico Rossetto

School of Computing Science
College of Science and Engineering
University of Glasgow



April 2026

Contents

1 Introduction **2**
- 1.1 Background and motivation . . . 2
- 1.2 Problem statement and research objectives . . . 5
  - 1.2.1 Thesis statement . . . 6
  - 1.2.2 Research questions . . . 6
- 1.3 Challenges in augmenting music representations with external text 7
  - 1.3.1 Coverage of textual knowledge . . . 7
  - 1.3.2 Reliability and consistency of textual knowledge . . . 8
  - 1.3.3 Granularity mismatch between textual and musical representations . . . 8
- 1.4 Experimental approach . . . 9
- 1.5 Contributions and key findings . . . 10
- 1.6 Thesis outline . . . 11

2 Related works **12**
- 2.1 Music representations from audio, and their blind spot . . . 13
- 2.2 Multimodal MIR: injecting semantics at inference time . . . 15
- 2.3 The current landscape: sources, pretraining, and evaluation . . . 16
  - 2.3.1 Textual knowledge sources for music . . . 16
  - 2.3.2 Learning from text: pretraining in single and multimodal setups . . . 18
  - 2.3.3 Evaluating music representations . . . 19
- 2.4 Research gaps . . . 21

3 Evaluating music representations **23**
- 3.1 Music representation framework . . . 25
  - 3.1.1 What are we evaluating? . . . 27
  - 3.1.2 Music representations vs model architecture . . . 28
- 3.2 Genre classification . . . 29
  - 3.2.1 Evaluation . . . 29
  - 3.2.2 Datasets . . . 30
- 3.3 Emotion recognition . . . 31
  - 3.3.1 Evaluation . . . 31
  - 3.3.2 Datasets . . . 32

3.4 Music tagging . . . 34
3.4.1 Evaluation . . . 34
3.4.2 Datasets . . . 36
3.5 Contributions . . . 37
3.6 Summary . . . 38

4 Artist biography corpora **39**
4.1 Augmenting music representations with external knowledge . . . 40
4.1.1 Why artist biographies . . . 41
4.2 Proposed corpora . . . 41
4.2.1 Web-scraped biographies . . . 41
4.2.2 LLM-based biographies . . . 47
4.3 Factual accuracy analysis . . . 51
4.3.1 Question answering evaluation . . . 51
4.3.2 Experimental results . . . 54
4.4 Contributions . . . 58
4.5 Summary . . . 58

5 Comparing music representations **60**
5.1 Evaluated representations . . . 61
5.1.1 Audio and metadata music representations . . . 61
5.1.2 Artist-biography augmented music representations . . . 61
5.1.3 Representation coverage . . . 62
5.1.4 Text-combined music representation . . . 64
5.2 Experimental setup . . . 65
5.2.1 Model architecture . . . 65
5.2.2 Training objectives . . . 68
5.2.3 Training and inference settings . . . 68
5.2.4 Evaluation protocol summary . . . 69
5.3 Performance of representations . . . 69
5.4 Multimodal fusion of audio and text representations . . . 72
5.4.1 Comparison with models from the literature . . . 73
5.5 Ablation studies . . . 75
5.5.1 Impact of the audio encoder . . . 75
5.5.2 Impact of artist-biography augmentations . . . 76
5.5.3 Masking in genre classification . . . 77
5.6 Contributions . . . 78
5.7 Summary . . . 79

6 Pretraining audio-text models **80**
6.1 Task reformulation . . . 81
6.1.1 Evaluation protocol . . . 81
6.1.2 Genre classification . . . 81
6.1.3 Emotion classification . . . 82
6.1.4 Scope relative to the benchmark suite . . . 83
6.2 Experimental setup . . . 84

6.2.1 Model architecture 84
6.2.2 Pretraining procedure 86
6.2.3 Fine-tuning procedure 88
6.3 Results 88
6.3.1 Distribution of music across the circumplex quadrants . . 90
6.3.2 The complexity of the reformulated tasks 92
6.4 Limitations and future work 93
6.4.1 Dataset overlap and artist supervision 93
6.4.2 Sources of artist biographies 94
6.4.3 Generative reformulation limitations 94
6.4.4 Scale of pretraining 95
6.5 Contributions 95
6.6 Summary 95

7 Discussion and conclusion 97
7.1 Summary of contributions 98
7.1.1 Evaluation methodology 98
7.1.2 Building the artist biography corpora 99
7.1.3 Multimodal music representations 99
7.1.4 Pretraining on artist biographies 99
7.2 Discussion 100
7.2.1 Revisiting the research hypotheses 100
7.2.2 Positioning within multimodal MIR research 101
7.2.3 Artist biographies as textual representations 101
7.2.4 Injection vs learning of textual knowledge 102
7.2.5 Interpretation of empirical findings 103
7.3 Limitations 104
7.3.1 Scalability 104
7.3.2 Genre leakage 104
7.3.3 Evaluation constraints 105
7.3.4 Dataset bias 105
7.4 Future work 106
7.4.1 Short-term directions 106
7.4.2 Longer-term perspectives 107
7.5 Research impact 108
7.6 Conclusion 109
7.7 Reflections 110

Abstract

Music Information Retrieval (MIR) aims to develop automated systems to analyse, categorise, and recommend music. Multimodal learning plays an important but challenging role in MIR. Compared to more common multimodal settings such as speech transcription or image captioning, aligning audio with textual descriptions of music is particularly difficult because many musical concepts are abstract, culturally grounded, and not directly observable in the signal.

Several approaches have explored incorporating textual information into MIR systems, leveraging lyrics, editorial metadata, or specialised audio-text corpora. However, these approaches exhibit a common trade-off between the semantic richness of the textual source and its availability and reproducibility, often due to coverage limitations or copyright restrictions.

In this thesis, we investigate artist biographies as an alternative source of textual supervision. We show that artist biographies provide a scalable and reproducible source of high-level semantic information. To this end, we construct two corpora: one derived from web-scraped, community-authored content, and one generated using a large language model. Both are made available for reuse, with a documented reconstruction procedure for the web-scraped text, whose licensing prevents direct redistribution. We further introduce a unified evaluation framework that enables controlled comparisons of representations across multiple semantic MIR tasks and allows us to assess the contributions of textual, audio, and multimodal inputs, both in isolation and in combination.

Experimental results show that incorporating artist biographies improves performance across most evaluated tasks. Relative to audio-only baselines, multimodal representations yield improvements of up to +38% in genre classification, +28% in emotion recognition, and +14% in music tagging, although one smaller dataset remains inconclusive due to limited training data. Additionally, we provide initial evidence that models can internalise the semantic information in biographies through generative pretraining, leading to improved performance even when textual inputs are unavailable at inference time.

These findings establish artist biographies as an effective, scalable, and reproducible source of semantic supervision for multimodal MIR and highlight the broader potential of structured textual knowledge for enhancing music representation learning.

Chapter 1

Introduction

Music Information Retrieval (MIR) is the field concerned with extracting, organising, and analysing information from music data. It encompasses a wide range of tasks that enable machines to understand and interact with music, including genre classification, emotion recognition, music tagging, similarity estimation, and music recommendation.

At its core, MIR seeks to bridge the gap between low-level audio signals and high-level semantic concepts used to describe music. While audio encodes acoustic properties such as timbre, rhythm, and harmony, many MIR tasks involve abstract and culturally grounded notions such as genre, mood, or style. These concepts are often subjective and context-dependent, making them difficult to infer directly from audio alone.

MIR systems play a central role in modern music applications, from streaming platforms and search engines to recommendation systems. Automatic tagging and classification enable large-scale organisation of music catalogues, similarity modelling supports playlist generation, and emotion recognition drives mood-based listening experiences. Figure 1.1 illustrates one such application: MoodAgent, a music streaming platform that generates playlists based on emotional dimensions such as *sensual*, *tender*, *happy*, or *angry*. As digital music collections continue to grow, the ability to automatically extract meaningful representations of music becomes increasingly important.

1.1 Background and motivation

In recent years, MIR has advanced substantially, driven by the availability of larger datasets, improved neural audio encoders [39], and the adoption of self-supervised pretraining approaches for music audio [29]. Alongside these developments, multimodal learning has emerged as a promising approach to modelling high-level semantic attributes of music by integrating audio representations with textual information [35]. Similar approaches have progressed rapidly in neighbouring fields such as vision and natural language processing, where models

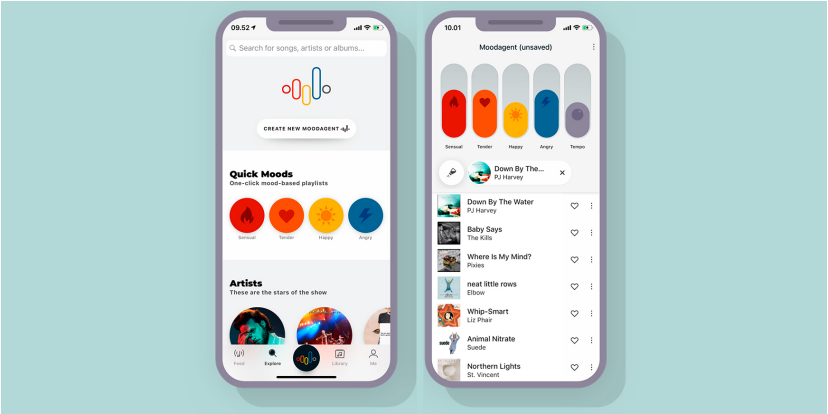


Figure 1.1: Screenshots of the MoodAgent music streaming application, illustrating mood-based playlist generation. The left panel shows a selection of mood categories (*Sensual*, *Tender*, *Happy*, *Angry*); the right panel shows a generated playlist with adjustable mood sliders. Such applications depend on MIR systems that can reliably associate audio content with semantic attributes like mood and emotion.

learn joint representations that link sensory signals to natural language descriptions. These developments have motivated increasing interest in integrating acoustic and textual information within MIR systems.

However, learning meaningful audio-text associations in music presents challenges that differ from those encountered in adjacent domains. In vision, text often describes directly observable entities such as objects, scenes, or actions, while in speech processing, text corresponds to transcribed linguistic content. In contrast, many of the concepts used to describe music (such as genre, mood, style, and descriptive tags) are culturally constructed and shaped by discourse surrounding artists and musical communities. These concepts are therefore not always deterministically recoverable from acoustic signals alone. As a result, audio-only models can struggle to capture high-level semantic attributes of music, particularly when supervision is sparse, labels are noisy, or stylistic categories overlap acoustically.

Existing approaches that incorporate textual information in MIR rely mainly on lyrics and on curated metadata or platform-derived annotations. Lyrics provide track-level language but are frequently unavailable for instrumental recordings and difficult to redistribute due to copyright restrictions. Metadata and platform annotations are often proprietary, task-specific, or not reproducible at scale. The field therefore lacks high-coverage, publicly available textual resources that can be systematically used for multimodal music representation learning.

To address these limitations, this thesis investigates artist biographies as an alternative source of semantic context for multimodal MIR. We first clarify the

key terms used throughout. A *track* is a single recorded piece of music, represented by an audio waveform and optionally accompanied by metadata such as title and artist name. An *artist* is the performing entity associated with one or more tracks and constitutes the organisational level at which biographies are defined. An *artist biography*, in the context of this thesis, is a short natural-language text that describes an artist's background, typically encoding information about musical style, genre, influences, geographic origin, career trajectory, and cultural context. Figure 1.2 illustrates a concrete example.

We also use the term *representation* in a specific sense that is central to this thesis. A *music representation* is not a learned embedding but the concrete input constructed for a music item and presented to the model: the transformation of the raw data (an audio waveform together with associated metadata) into the signal the model consumes, such as a mel-spectrogram derived from the audio or a textual description templated from metadata and artist biographies. A single track can therefore be rendered into several different representations, and the same music item can be encoded through different *representation strategies*: audio alone, a metadata-derived text, or that text further augmented with a biography. This raw-data-to-input transformation is the object of study throughout. Because the downstream model and training protocol are held fixed, differences in performance can be attributed to the representation itself rather than to model capacity, which is what makes the representation, rather than the encoder, the quantity worth naming and comparing. Chapter 3 formalises this notion and sets out the controlled protocol used to evaluate the quality of a representation.

Throughout this work, we use the term *textual knowledge* to refer broadly to any natural-language source that provides semantic information about a music item or its associated artist, such as lyrics, editorial metadata, community-authored descriptions, or biographies. Within this category we distinguish *structured* from *unstructured* textual knowledge. The distinction is not one of format, since both are written in free-flowing natural language and neither follows a formal schema or fixed field layout. It is one of *predictability of content*: whether the text can be expected to encode a recurring set of semantic categories across most items.

Text is *structured* in this sense when it consistently encodes recognisable categories of information, such as genre, geographic origin, artistic influences, and career history, so that for a given artist one can reasonably anticipate which kinds of attributes will be present. Artist biographies are the primary example: they vary in style and length, but reliably convey these recurring attributes, which is what makes them usable as a systematic source of supervision. *Unstructured* textual knowledge, such as social-media posts, forum discussions, promotional taglines, or user comments, may mention an artist or track without dependably encoding any of these categories, and its informativeness varies unpredictably from one item to the next. The boundary is a matter of degree rather than a hard line, and biographies sit firmly on the structured side of it.

Artist biographies encode relatively stable information about musical style, influences, geography, and artistic identity, offering contextual signals that com-

> *"Marshall Bruce Mathers III (born October 17, 1972), known professionally as Eminem (often stylised as* EMINƎM*), is an American rapper, songwriter, record producer, record executive, film producer, and actor. He is consistently cited as one of the greatest and most influential rappers of all time and has been labelled the "King of Hip Hop. In addition to his solo career, Eminem was a member of the hip hop group D12. He is also known for his collaborations with fellow Detroit-based rapper Royce da 5'9"; the two are collectively known as Bad Meets Evil. After his debut album "Infinite" (1996) and the extended play "Slim Shady EP" (1997), Eminem signed with Dr. Dre's Aftermath Entertainment and subsequently achieved mainstream popularity in 1999 with "The Slim Shady LP", which earned him his first Grammy Award for Best Rap Album."*

Figure 1.2: **Example of an artist biography** for Eminem, sourced from Wikipedia via the KILT benchmark [38]. In a single paragraph, it encodes geographic origin (*American*), genre (*rapper*, *hip-hop*), cultural impact (*one of the greatest and most influential rappers of all time*), career history (*signed with Dr. Dre's Aftermath Entertainment*), and collaborative context (*member of the hip hop group D12*). These are the kinds of structured semantic cues that this thesis argues can complement acoustic representations.

plement acoustic representations. Importantly, such biographies can be collected or generated in a reproducible manner, enabling systematic experimentation at scale.

This thesis, therefore, explores whether artist-level textual knowledge can improve music representation learning. It constructs and evaluates two collections of artist biographies, analyses their reliability and coverage, and investigates their usefulness across several semantic MIR tasks, including genre classification, emotion recognition, and music tagging. In addition, the thesis examines whether models can incorporate information from biographies in two complementary ways: through explicit multimodal augmentation during inference and through generative pretraining that encourages audio models to internalise contextual textual knowledge.

1.2 Problem statement and research objectives

Multimodal learning in Music Information Retrieval (MIR) offers a promising approach to modelling not only the acoustic characteristics of music but also the semantic and contextual dimensions that shape how music is described, categorised, and interpreted. Many MIR tasks, among them genre classification, emotion recognition, and music tagging, reflect culturally grounded and stylistic concepts that are not always recoverable from audio signals alone. Incorporating textual information can therefore provide complementary semantic cues.

However, the development of multimodal MIR systems is constrained by the limited availability of high-coverage, reproducible textual resources. Existing approaches typically rely on lyrics, expert-curated metadata, or proprietary platform annotations. Lyrics are frequently unavailable for instrumental tracks and are subject to copyright restrictions. Expert-created corpora require substantial manual effort and are often task-specific or inaccessible for redistribution. What is missing, then, is scalable and publicly available textual supervision whose impact on music representation learning can be systematically evaluated.

Furthermore, despite growing interest in multimodal architectures, there remains a limited understanding of how structured textual knowledge influences learned representations across different MIR tasks under controlled experimental conditions. This gap motivates the need for both accessible textual resources and rigorous evaluation protocols.

1.2.1 Thesis statement

This thesis demonstrates that structured artist-level biographies provide an effective and reproducible source of semantic supervision for music representation learning. It shows that integrating biography-derived textual knowledge with audio representations improves performance across genre classification, emotion recognition, and music tagging, with relative gains of up to +38%, +28%, and +14%, respectively, over audio-only baselines across six benchmark datasets. While results on one smaller dataset remain inconclusive, the findings are otherwise consistent. Separately, ablation experiments that mask explicit genre mentions in the biographies show that the genre gains are not reducible to label leakage. Furthermore, the thesis provides evidence that generative pretraining on artist biographies enables models to internalise contextual knowledge, improving downstream performance even when textual input is unavailable at inference time.

1.2.2 Research questions

Answering questions about the value of a representation requires controlled experimental conditions: a framework in which representations can be compared while the downstream model and training protocol are held fixed. Establishing such a framework is itself a methodological contribution of this thesis rather than an open empirical question: it is the instrument that makes the questions below answerable, and it is developed in Chapter 3. The thesis is organised around three research questions, each paired with an explicit hypothesis:

RQ1 (Resource). Can artist biographies be constructed at scale to yield high-coverage, reliable, and reproducible textual knowledge for music catalogues? We hypothesise that, unlike lyrics or curated metadata, artist-level information is abundant enough online to be aggregated through web scraping and synthesised with large language models, producing corpora that cover the majority of artists in standard MIR datasets.

RQ2 (Multimodal effect). When artist biographies are fused with audio in a multimodal model at inference time, do they improve performance on semantic MIR tasks relative to audio-only baselines, and how does their contribution compare with that of existing textual sources such as metadata and lyrics? We hypothesise that biographies provide complementary semantic context that improves performance, with the largest gains on tasks that depend on stylistic and cultural attributes rather than on purely acoustic ones.

RQ3 (Internalisation). Can the contextual knowledge contained in artist biographies be internalised by an audio model through generative pretraining, such that it improves downstream performance even when no text is available at inference time? We hypothesise that biography-driven pretraining yields improvements over direct fine-tuning, but that these gains are smaller than those obtained through explicit multimodal augmentation, because internalisation requires the model to compress contextual knowledge into its parameters rather than accessing it directly.

The three questions are distinct in kind. RQ1 concerns the *resource* itself: whether reliable biography corpora can be built at all, independent of any downstream use. RQ2 concerns the *utility* of that resource when it is supplied to a model at inference time. RQ3 concerns whether the same knowledge can instead be *internalised* during training, so that no text is needed at inference. Together they trace a single line of inquiry from constructing the resource, to using it explicitly, to absorbing it implicitly.

1.3 Challenges in augmenting music representations with external text

Augmenting music representations with textual knowledge poses challenges in aligning textual information with musical content. While text provides valuable semantic context that may complement acoustic representations, the relationship between textual discourse and musical signals is indirect and often uneven. In practice, integrating textual knowledge into Music Information Retrieval (MIR) systems requires addressing issues of coverage, reliability, and representational granularity. These three challenges map directly onto the research questions above: coverage and reliability determine whether a usable resource can be built at all, and therefore frame RQ1, while the granularity mismatch shapes how effectively that resource can be exploited once built, whether through explicit fusion at inference (RQ2) or through internalisation during pretraining (RQ3).

1.3.1 Coverage of textual knowledge

A first challenge concerns the availability of textual information associated with music recordings. Many textual resources used in multimodal MIR, such as

lyrics or curated metadata, suffer from incomplete coverage. Lyrics are unavailable for instrumental recordings and are often subject to copyright restrictions, limiting their accessibility for reproducible research. Encyclopedic resources such as Wikipedia provide detailed information on well-known artists but often underrepresent emerging or independent musicians. Similarly, community-authored content may cover niche musical scenes while remaining sparse or inconsistent across broader catalogues.

As a result, the presence of textual information varies significantly across datasets and artists. When textual augmentation is used to enrich music representations, this uneven availability introduces the risk that performance differences reflect the presence or absence of text rather than the intrinsic usefulness of the representation itself. Coverage is thus a design consideration in its own right when building and evaluating knowledge-augmented music representations. This challenge motivates RQ1 directly: a usable biography corpus must not only exist but achieve demonstrably high coverage across the datasets on which representations are evaluated.

1.3.2 Reliability and consistency of textual knowledge

Beyond coverage, textual resources vary widely in terms of reliability and consistency. Artist descriptions may originate from heterogeneous sources, including encyclopedic entries, promotional biographies, community-written profiles, or automatically generated text. These sources differ substantially in style, structure, and factual accuracy. Some descriptions emphasise stylistic narratives or marketing language rather than verifiable information, while automatically generated content may introduce incorrect or hallucinated facts.

This variability makes it difficult to assume that textual resources consistently provide reliable representations of musical context. When textual information is used as an input to representation learning, inaccuracies or inconsistencies in the underlying text may propagate into the learned representations. Assessing the quality and reliability of textual knowledge is therefore integral to constructing corpora intended to augment music representations. This concern forms the second half of RQ1: a biography corpus must be judged not only on how broadly it covers a catalogue but on how reliably its content reflects verifiable artist information, a property that Chapter 4 evaluates explicitly through a Wikidata-grounded protocol.

1.3.3 Granularity mismatch between textual and musical representations

A further challenge arises from differences in the level of abstraction at which textual and musical information are described. Audio representations typically operate at the level of individual recordings, capturing acoustic properties specific to a particular track. In contrast, many textual descriptions of music operate at broader levels of abstraction. Artist biographies, for example, often describe stylistic influences, career trajectories, and cultural context associated

with a musician rather than the properties of a specific recording. Other approaches have explored textual resources, such as album reviews, which describe music at the level of an album rather than individual tracks or artists [35].

This difference introduces a granularity mismatch between textual and acoustic representations. Artist-level descriptions may provide stable contextual information about stylistic tendencies or musical identity, but they may also overlook variation between individual tracks. More broadly, textual discourse about music frequently reflects cultural interpretation and categorisation rather than direct descriptions of acoustic structure. Aligning such contextual information with track-level audio representations, therefore, requires careful consideration of how textual knowledge is incorporated into music representation learning. This challenge underlies RQ2 and RQ3, which examine two contrasting ways of bridging the granularity gap: supplying artist-level text alongside track-level audio through explicit multimodal fusion at inference (RQ2), and encouraging the model to internalise that contextual knowledge during pretraining so that it informs track-level predictions without text at inference (RQ3).

1.4 Experimental approach

To investigate these challenges and evaluate the potential of artist biographies as a source of semantic supervision, this thesis adopts a structured experimental methodology designed to isolate the impact of textual knowledge on music representation learning.

First, a controlled evaluation framework is established to assess music representations across multiple semantic MIR tasks. Within this framework, the model architecture and evaluation protocol are held constant to isolate the contribution of the representation itself. This design enables systematic comparison of different representation strategies (audio, textual, and multimodal) while limiting confounding factors introduced by architectural variation. The evaluation pipeline spans three tasks commonly used to probe semantic understanding in MIR: genre classification, emotion recognition, and music tagging, providing complementary perspectives on the semantic information captured by a representation.

Second, the thesis constructs two complementary corpora of artist biographies designed to provide broad coverage while remaining reproducible and, within the limits of the underlying content licences, openly accessible. These corpora are derived from community-authored sources and large language model-generated text, allowing the study to balance coverage and stylistic consistency. Their properties are analysed with respect to coverage and factual reliability to assess their suitability as sources of textual supervision for music representation learning.

Third, the impact of textual augmentation is evaluated by comparing audio-only, textual, and multimodal representations within the proposed evaluation framework. These experiments examine whether artist biographies provide semantic information that complements acoustic representations and whether such

information can be leveraged by multimodal models to improve performance on downstream MIR tasks.

Finally, the thesis investigates whether contextual textual knowledge can be internalised within an audio model through generative pretraining. Instead of providing textual information at inference time, an audio-text model is pretrained to reconstruct textual representations from audio. This approach allows the model to learn contextual associations between acoustic patterns and structured textual knowledge during training, enabling the study of whether such knowledge can influence downstream performance even when textual input is unavailable.

1.5 Contributions and key findings

This thesis investigates how structured textual knowledge can enhance music representation learning for semantic Music Information Retrieval tasks. Through a combination of controlled evaluation, corpus construction, multimodal experimentation, and generative pretraining, the work examines both the practical and methodological implications of incorporating artist-level textual information into MIR systems. The main contributions of this thesis are as follows:

- **A controlled evaluation framework for music representation learning.** The thesis establishes a unified evaluation pipeline designed to isolate the effect of representation choice on downstream MIR tasks. By fixing model architectures, dataset splits, and evaluation protocols across experiments, the framework enables systematic comparison of audio-only, textual, and multimodal representations across genre classification, emotion recognition, and music tagging tasks.
- **Two reproducible corpora of artist biographies for multimodal MIR research.** To address limitations in the availability of textual resources, the thesis introduces two complementary collections of artist biographies derived from community-authored sources and large language model-generated text. These corpora provide high coverage across multiple MIR datasets while remaining fully reproducible. Their properties are analysed with respect to coverage and factual reliability using a Wikidata-grounded evaluation protocol.
- **A systematic comparison of audio, textual, and multimodal music representations.** Using the proposed evaluation framework, the thesis demonstrates that incorporating artist biographies improves performance across several MIR tasks when combined with audio representations. The experiments show that artist-level textual information provides complementary semantic cues that can enhance models' ability to capture stylistic and contextual attributes of music.
- **Initial evidence that contextual textual knowledge can be internalised through generative pretraining.** Beyond explicit multimodal fusion, the thesis explores whether artist-level textual knowledge can shape audio representations during training. By pretraining an audio-text model

to reconstruct biography-based textual representations from audio, the study provides initial evidence that contextual knowledge can be internalised and improve downstream performance even when textual inputs are unavailable at inference time.

1.6 Thesis outline

The remainder of this thesis is organised as follows.

Chapter 2 reviews related work in Music Information Retrieval and multimodal learning. It surveys approaches to audio-based music representation learning and prior work that incorporates textual information, including lyrics, metadata, and other forms of contextual knowledge. This chapter situates the thesis within the broader landscape of multimodal MIR research and highlights the limitations of existing textual resources.

Chapter 3 introduces the evaluation framework used throughout the thesis. It formalises the concept of music representations and presents a controlled experimental pipeline for assessing the quality of representations across multiple semantic MIR tasks. The chapter defines the selected tasks, datasets, and evaluation metrics, establishing a unified benchmark for the experiments that follow.

Chapter 4 presents the construction of two corpora of artist biographies designed to provide high coverage while remaining reproducible. The chapter describes the data collection and generation processes, analyses the properties of the resulting corpora, and evaluates their factual reliability using a Wikidata-grounded question answering framework. The presented corpora are designed to be reproducible and are made available to support future research on multimodal music representation learning, with a documented reconstruction procedure for the web-scraped text, whose licensing prevents direct redistribution.

Chapter 5 investigates the impact of textual augmentation on music representation learning. Audio, textual, and multimodal representations are systematically compared across the tasks of genre classification, emotion recognition, and music tagging. The experiments analyse the complementary contributions of acoustic and textual information and examine how different textual sources influence model performance.

Chapter 6 explores whether contextual textual knowledge can be internalised through generative pretraining. An audio-text model is pretrained to reconstruct artist-level textual representations from audio and subsequently fine-tuned on downstream tasks. The chapter evaluates whether such biography-driven supervision improves performance when textual input is unavailable at inference time.

Chapter 7 concludes the thesis by discussing the implications of the findings, outlining limitations of the current work, and identifying directions for future research on knowledge-aware music representation learning.

Chapter 2

Related works

Chapter 1 argued that many of the concepts music information retrieval (MIR) cares about, among them genre, mood, style, and the tags listeners use to organise their libraries, are culturally constructed and only weakly recoverable from the audio signal alone. The central claim of this thesis follows from that observation: if such concepts live in the discourse *around* music rather than solely in the waveform, then a textual source that captures that discourse should improve how well a model represents music, provided the source is rich enough to be informative and yet scalable and reproducible enough to be used and shared. This thesis proposes *artist biographies* as such a source. The purpose of this chapter is to substantiate that claim by tracing the line of work that leads to it, and to show precisely where the existing literature stops short.

This thesis is concerned with three families of semantic MIR tasks that recur throughout the chapter and the remainder of the document: *genre classification*, *emotion recognition*, and *music tagging*. These are the tasks on which the representations studied here are evaluated, and they are introduced informally as they arise below; their datasets and metrics are consolidated in Section 2.3 and formalised in Chapter 3. Together they probe complementary facets of musical understanding: discrete stylistic categories, graded affective content, and open-vocabulary descriptive labels.

Scope and approach of this review. This chapter is a focused rather than exhaustive review. Rather than survey the whole of MIR, it follows the argument of the thesis and reviews the work that bears directly on four questions, taken up across the three sections below: how music is represented from audio, how additional modalities have been combined with audio, what textual resources and pretraining objectives the field has used to supply semantic knowledge, and how such representations are evaluated. The literature was gathered primarily from the core MIR venues (ISMIR and its associated journal TISMIR, IEEE TASLP, and ICASSP) and, for cross-modal representation-learning methods that originate outside MIR, from the main machine-learning and language venues (NeurIPS, ICML, and ACL). Works were selected by combining seminal

references with representative and recent examples of each line of work, and by expanding citations outward from the small number of papers that combine audio with text for music. Throughout, the organising lens is *representation-centric*: work is discussed in terms of the information source and learning signal that shape a representation, rather than in terms of specific downstream architectures, consistent with the framing developed in Chapter 3. The experimental work of this thesis was carried out in 2023, and the review is bounded accordingly: it reflects the state of the field up to that point. Developments since then are noted only where they bear on the interpretation of the contributions (Section 2.3.2), rather than surveyed in full.

The chapter is organised as a single argument in three parts, closing with the gaps it exposes. Section 2.1 reviews music representations learned from audio alone and identifies their recurring blind spot: high-level, culturally grounded context. Section 2.2 reviews multimodal MIR, which addresses that blind spot by injecting non-audio modalities at inference time, and asks what such systems have and have not established. Section 2.3 surveys the current landscape that any new proposal must situate itself within: the available textual knowledge sources, the pretraining objectives that internalise text during training, and the frameworks used to evaluate representations. Finally, Section 2.4 consolidates the three gaps that emerge, concerning the *source* of textual knowledge, the *control* of comparisons across it, and the *stage* at which it is used, and maps them onto the research questions of this thesis.

2.1 Music representations from audio, and their blind spot

Audio is the primary modality in MIR, and for most of the field's history the question of how to represent music has been the question of how to turn a waveform into features a model can use. The dominant practice is to convert the signal into a time-frequency representation and then map it into a compact embedding. Early systems relied on hand-engineered timbral descriptors, most prominently *Mel-frequency cepstral coefficients* (MFCCs), a compact summary of the short-term power spectrum on the perceptually motivated Mel scale, originally developed for speech and adopted wholesale for music [31]. Paired with shallow classifiers, MFCCs established the first widely used baselines for tasks such as genre recognition [49]. Their limitation is instructive: because they discard most temporal and harmonic structure, they capture how music sounds locally but little of how it is organised, which set the agenda for representations learned directly from data.

The deep-learning era replaced hand-crafted descriptors with representations learned from the *log-mel spectrogram*, a time-frequency image that preserves the harmonic and timbral structure MFCCs discard while remaining compact enough to train on. This front-end has become the de facto standard input to convolutional audio models in MIR, and it is the representation adopted

throughout this thesis (Chapter 3). Large-scale supervised classification established convolutional networks as general-purpose audio feature extractors: VGGish, trained on the AudioSet corpus, is the canonical example and remains a widely used off-the-shelf audio front-end [21, 19]. Within music specifically, early work showed that convolutional representations learned end-to-end from tagging objectives were a practical alternative to feature engineering for recommendation and auto-tagging [14, 39], that combining convolutional feature extraction with temporal modelling suited music tagging and classification [10, 53], and that aggregating pretrained features across receptive fields gave strong transfer to downstream tasks [27, 11]. The common thread is that a supervised objective at scale yields an embedding that transfers, motivating the treatment of a learned audio encoder as a reusable representation layer rather than a task-specific component.

More recent work has shifted from supervised tagging toward self-supervised learning (SSL), on the premise that the structure and invariances of music can be learned from large unlabelled corpora without the cost and bias of annotation. Transformer-based and contrastive SSL encoders now dominate general-purpose music representation, from Swin-transformer pretraining for classification [55] to contrastive learning of musical representations [46] and masked-prediction encoders such as MERT and related models [29, 28], echoing the trajectory of speech representation learning [3, 22]. These approaches matter to this thesis as a point of reference: they represent how far audio-only pretraining can go before external information is introduced, and syntheses of supervised and self-supervised music understanding confirm that both routes yield strong, transferable audio embeddings [33].

Yet strong audio embeddings share a blind spot that no amount of audio-only pretraining removes. A representation learned from the signal can encode what a piece *sounds like*, but not the contextual knowledge that determines how listeners categorise it: the lineage of a genre, an artist's influences and trajectory, the scene or era a record belongs to, or the cultural associations a style carries. This applies across the semantic tasks considered in this thesis. Genre categories encode lineage and scene as much as timbre, and the descriptive tags used in auto-tagging mix acoustic attributes with contextual and evaluative ones, so both rest partly on conventions about how music is described rather than on the signal alone. Affective MIR illustrates the point with particular clarity, since a substantial body of work models emotion not as a categorical label but as position in a continuous valence–arousal space [44, 51], sometimes as a temporally evolving trajectory [52], and the resulting targets depend on interpretive and cultural cues only partially present in the acoustics. Genre, mood, and descriptive tags are, in this sense, judgements about music as much as properties of the signal. It is precisely this gap between acoustic content and contextual meaning that motivates supplementing audio with an external, textual source of semantics, which is the subject of the rest of the chapter.

2.2 Multimodal MIR: injecting semantics at inference time

The natural response to the blind spot of audio-only models is to give the model access to another modality that carries the missing semantics. Multimodal MIR does exactly this, pairing audio with text, images, or metadata so that representations are shaped by contextual signals as well as acoustic ones. A consistent finding across this literature is that adding a non-audio modality improves performance on semantically loaded tasks relative to audio-only baselines, and that the gains are largest precisely where audio is weakest.

The clearest evidence comes from genre classification, where combining audio with text and cover-art imagery yields representations that outperform any single modality [35, 36]. The improvement is generally attributed to *complementarity*: the non-audio modalities supply discriminative information, such as editorial descriptions, reviews, and visual style cues, that the audio branch does not recover on its own. These signals are also frequently available at the artist or album level rather than per track, so multimodal systems must integrate information across granularities. Both properties recur throughout this thesis, the second in particular foreshadowing the artist-level supervision it relies on. A parallel line of work reaches the same conclusion for emotion: models that fuse audio with lyrics outperform unimodal baselines on mood and emotion recognition, because lyrical content carries affective cues that the signal expresses only weakly [13], and cross-modal attention architectures have been developed to model the interaction between the modalities explicitly [56]. Beyond curated pairings, the broader multimodal-learning literature has shown that pretrained cross-modal spaces can be repurposed for music with little or no task-specific supervision [8].

In all of these systems, text (or another modality) is used in the same way: as an *additional input available at inference time*. The model receives audio and text together and fuses them to make a prediction; remove the text at test time and the benefit disappears. This inference-time use, which the thesis later terms the *injection* of textual knowledge, is one of two complementary ways text can enter a music model, and it is the mode studied directly in Chapter 5.

What this literature establishes is that multimodality helps. What it leaves unresolved is *why*, and *under what conditions*. Reported gains are obtained with different text sources (reviews, lyrics, metadata, tags), different fusion architectures, and different datasets and splits, so it is rarely possible to tell whether an improvement comes from the information in the text, the capacity added by a second encoder, or the particular benchmark used. In particular, the field seldom isolates the effect of *which* text is used while holding everything else fixed, leaving open whether richer, more contextual text yields measurably larger gains than short, label-like text, or whether adding any text at all is what matters. Answering that question requires both a controlled way of comparing text sources and a clear view of what sources are available, which brings the review to the current landscape.

2.3 The current landscape: sources, pretraining, and evaluation

Situating a new textual source within multimodal MIR requires an account of three things: what textual resources the field already has and what they lack, how text has been used not only at inference but as a pretraining signal, and how representations built from these ingredients are evaluated. This section reviews each in turn, and each exposes one of the gaps the thesis addresses.

2.3.1 Textual knowledge sources for music

Textual sources for music vary enormously in how much they say and how easily they can be obtained, and the range is worth seeing concretely before considering the trade-off they collectively impose. At one extreme sit *social tags*, short keyword-like annotations such as `chill`, `90s`, or `female vocalist`, which are abundant and cheap but encode only sparse, categorical hints. *Editorial metadata* is similarly compact: a typical track record amounts to little more than structured fields such as `Author: The Trammps` and `Title: Disco Inferno`. A *promotional blurb* or scraped web description adds a sentence or two of free text but is often noisy and inconsistent; a real community-authored entry may read only "*A Way Of Life, A Collective of Hip-Hop from NJ*". At the far extreme, a full *artist biography* narrates influences, geography, career, and style: "*The Trammps were an American disco and soul group from Philadelphia, Pennsylvania. Formed in the early 1970s ... known for their energetic and danceable music ...*". These examples make concrete a distinction the thesis relies on: sources differ not only in length but in the *predictability* of what they encode, with biographies reliably conveying recurring stylistic and contextual categories that tags and metadata do not.

Each source has been studied as a supervision signal, and each carries a characteristic limitation. *Lyrics* are the most direct track-level textual signal and are effective for mood and emotion [13], but coverage is incomplete (instrumental music has none, and transcriptions are missing or noisy) and licensing constraints complicate redistribution and thus reproducibility. *Editorial and community-curated metadata* is exemplified by representation learning from Discogs (a large community-maintained database of music releases and their editorial genre, style, and credit annotations), which encodes semantic structure that is not recoverable from audio and generalises across an artist's catalogue [2]; but it is platform-specific, biased toward particular markets, and rich in categorical labels while poor in narrative context. *Free-form web text* (reviews, biographies, descriptions) is semantically richer and remains available when lyrics are not, but it is noisy, heterogeneous, hard to curate at scale, and legally fraught to redistribute. A common way to exploit such text despite its noise is to treat it as *weak supervision*, pairing audio with rough natural-language descriptions [32]; tellingly, even here the large audio-text corpora that make the approach work are often not publicly redistributable, which limits

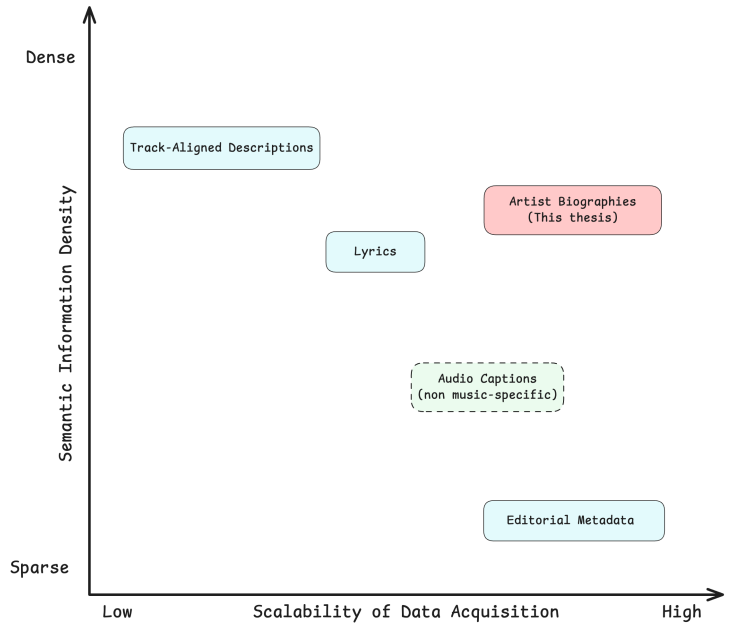


Figure 2.1: Qualitative positioning of textual knowledge sources used in MIR along two axes: semantic information density and scalability of data acquisition. Track-aligned descriptions [32, 1] and lyrics [13] offer high semantic richness but are constrained in scalability by production cost and licensing, respectively. Audio captions [17] are scalable but not music-specific, limiting their semantic density for MIR tasks. Editorial metadata [2] is highly scalable but encodes sparse, categorical information. Artist biographies (this thesis) occupy a distinctive position: combining high semantic density with broad scalability, enabled by community-authored and LLM-generated sources.

reproducibility and controlled comparison.

Artist-level text specifically has been used before, which is important to acknowledge so that the contribution of this thesis is stated precisely. Artist biographies have been used to learn artist embeddings for cold-start music recommendation, combining biographical text with audio and usage data [34]; and Wikipedia is the canonical large-scale repository of entity-level biographical knowledge, standardised for retrieval by benchmarks such as KILT [38]. These resources demonstrate that artist-level knowledge is useful and available, but they also reveal what is missing: existing artist-text resources are either tied to a specific application and proprietary usage data [34], or general-purpose encyclopaedic corpora not curated for music and skewed toward well-documented artists [38]. None provides a music-focused, high-coverage, openly redistributable corpus of artist biographies suited to controlled MIR experiments.

Taken together, these sources trace out a single trade-off, summarised in Figure 2.1: semantic richness tends to come at the cost of scalability and reproducibility, and vice versa. Lyrics and track-aligned descriptions are rich but hard to obtain at scale or to redistribute; tags and editorial metadata are scalable but sparse; general audio captions are scalable but not music-specific. The upper-right region, combining high semantic density with broad, redistributable scale, has remained largely unoccupied. This is the first gap the thesis exploits: artist biographies are argued to sit in that region, being narratively rich like web text yet, when deliberately constructed, coverable across a catalogue and distributable under clear terms. Constructing and validating such a corpus is not a matter of prior work but a contribution of this thesis, and is the subject of Chapter 4.

2.3.2 Learning from text: pretraining in single and multimodal setups

The multimodal systems of Section 2.2 use text at inference. A second, complementary way to use text is during *pretraining*, where language shapes the audio representation itself and need not be present downstream. Understanding this distinction is essential, because it is the axis along which the two experimental parts of this thesis divide.

The most influential pretraining recipe aligns modalities contrastively. CLIP established the template in vision, training paired image and text encoders so that matching pairs are close in a shared space [40]; the idea transferred to audio through AudioCLIP [20] and, at scale, CLAP [17], and to music specifically through MuLan, which aligns music audio with unconstrained natural-language descriptions to support retrieval and zero-shot classification [23]. Across these systems, text acts as a semantic anchor that reshapes the audio representation through cross-modal alignment. A complementary strand uses *generative* objectives inspired by language modelling: codified audio language modelling learns transferable representations by predicting audio token sequences [7], while text-conditioned generation systems such as MusicLM show that language can serve as a high-level control and supervision signal for music [1]. Weak language supervision [32] can be seen as an early, less structured member of this family.

Two observations about this literature motivate the thesis. Whether contrastive or generative, the textual supervision is almost always *short and surface-level*, consisting of tags, captions, prompts, or brief descriptions rather than rich, entity-centric knowledge about artists, so the richer target of artist biographies remains largely unexplored as a pretraining signal. More fundamentally, the field treats the two uses of text, injection at inference and internalisation during pretraining, largely in isolation. It studies one or the other rather than asking whether knowledge supplied as an explicit modality can instead be *absorbed* into the audio encoder so that its benefit survives when text is removed. This matters in practice: text is often unavailable or undesirable at inference, and a benefit realised only when text is present is a benefit that many deployments cannot use. Whether artist knowledge can be internalised through a generative

pretraining objective, by pretraining an audio-to-text model to produce artist biographies from audio and then fine-tuning on audio-only tasks, is the question taken up in Chapter 6, and it complements rather than replaces strong audio-only pretraining such as CLMR and MERT [46, 29].

Since this work was carried out, the audio-language direction has advanced considerably. Large language models have been used to synthesise music-caption corpora at scale [15], and instruction-tuned music-language models now perform open-ended music understanding, captioning, and question answering [18, 30]. These systems postdate the experiments reported here and are therefore not part of the comparison; they are noted because they reinforce rather than supersede the premise of this thesis. Each still depends on rich natural-language descriptions of music, and each inherits the same open questions about where such text is sourced, whether it can be shared and reproduced, and how the contribution of a given source can be isolated, so the gaps identified below apply to them as directly as to the earlier work reviewed above.

2.3.3 Evaluating music representations

Whatever the source of a representation, a claim that it is better than another is only as credible as the protocol used to compare them. MIR has a long tradition of evaluation, beginning with shared-task campaigns such as MIREX and the broader IMIRSEL effort, which established common task formulations, metrics, and reporting practices [16]. This tradition was a genuine advance over ad hoc dataset construction, but it is insufficient for the questions this thesis asks, for concrete reasons. Its evaluation-by-submission model reported system rankings without releasing the underlying data or, in many cases, the systems, so results could be compared but not independently reproduced or re-analysed; task definitions and datasets evolved between editions, undermining longitudinal comparison; and the emphasis fell on end-to-end *systems* for a task rather than on the reusable *representations* that are the object of study here.

A second, methodological literature explains why these constraints are not merely inconvenient. Analyses of MIR evaluation show that apparent rankings of methods are highly sensitive to design choices such as dataset sampling, annotation noise, evaluation splits, and statistical testing, and, critically for representation learning, to confounds such as artist overlap between training and test sets, production-era cues, and recording conditions that let a model exploit spurious correlations instead of the intended semantics [50]. The risk is heightened for multimodal systems, where a non-audio signal may encode identity or catalogue information that leaks across splits. Robust conclusions therefore demand careful split design, transparent reporting, and, when comparing text sources, explicit control for what each source might inadvertently reveal.

In response, representation learning has motivated benchmarks that evaluate embeddings directly as reusable features. HEAR (the Holistic Evaluation of Audio Representations) is the clearest example beyond MIR: rather than a single dataset, it is a standardised pipeline that freezes a candidate representation and probes it with lightweight classifiers across a diverse battery of downstream au-

Dataset	Task	Availability	Principal limitations
GTZAN [49]	Genre	Public audio	Small; documented faults: duplicate clips, mislabellings, and artist replication that leaks identity across splits [26]
FMA-Small [12]	Genre	Public, Creative Commons	Coarse genre boundaries; skewed toward long-tail independent artists
EmoMusic [45]	Emotion	Public audio	Very small and low-resource; continuous valence–arousal annotations are noisy and costly
Deezer [13]	Emotion	Annotations public, **audio proprietary**	Audio not redistributable, undercutting reproducibility; affective targets derived automatically
MagnaTagATune [25]	Tagging	Public audio	Tag vocabulary noisy, redundant/synonymous, and long-tailed
Jamendo-MoodTheme [6]	Tagging	Public, Creative Commons	Subjective mood/theme labels; strong class imbalance

Table 2.1: **Survey of the MIR benchmark datasets relevant to this thesis, emphasising availability and limitations.** The datasets span the three task families and a range of annotation regimes, but differ sharply in redistributability and in the confounds they introduce. Chapter 3 adopts this set and details the scales, metrics, and splits used for each.

dio tasks, including music, so that comparison reflects the representation rather than a bespoke downstream model [48]. Within MIR, MARBLE pursues the same goal, offering a unified protocol for benchmarking music representations across a taxonomy of tasks with baseline implementations, while foregrounding the practical obstacles of dataset access and copyright that determine whether such evaluation can be reproduced at all [54]. These frameworks are conceptually aligned with this thesis: they evaluate *transferability* across many tasks and reduce the chance that a conclusion is an artefact of one downstream model.

Because this thesis compares representations empirically, it inherits both the tasks and the datasets on which the field conducts these evaluations. Table 2.1 surveys the six benchmarks that recur across genre classification, emotion recognition, and music tagging, asking of each what it can and cannot support: how freely it can be redistributed, and which confounds it introduces. Public audio exists for all but Deezer, whose reliance on proprietary audio directly undercuts reproducibility, while the smaller sets (GTZAN, EmoMusic) and the noisier tag vocabularies (MagnaTagATune) bound what can be concluded from any single benchmark. This is a survey of the available resources and their limits, and it motivates the reproducibility concerns that run through this thesis. Chapter 3 takes the same six datasets and asks a different question of them, namely what a controlled evaluation protocol built on them looks like, and tabulates the tasks, label spaces, and metrics accordingly.

The unifying lesson of this evaluation literature is a gap rather than a solution. Across multimodal MIR, comparisons in the literature are frequently confounded by differences in model architecture, datasets, and evaluation protocol,

so the field has relatively few studies that isolate the effect of *which* representation, and for this thesis which *text source*, is used while holding downstream capacity, fusion strategy, and splits constant. Establishing such a controlled comparison is a prerequisite for the empirical claims of this thesis and is the purpose of the framework developed in Chapter 3.

2.4 Research gaps

The three sections of this chapter converge on three gaps, which together define the research programme of the thesis. Each has been raised as an observation about the literature above; here they are stated as the openings the thesis sets out to fill, and mapped to its research questions and contribution chapters.

The first gap concerns the *source* of textual knowledge. Text has repeatedly been shown to complement audio for semantic MIR tasks, through weak language supervision [32], editorial metadata [2], multimodal genre systems [35, 36], and audio-lyrics models for emotion [13], yet the sources commonly used sit at the extremes of the richness-scalability trade-off of Figure 2.1. Lyrics are track-aligned but sparse and licence-encumbered; tags and metadata are scalable but shallow; large web and platform corpora are rich but proprietary, unstable, and hard to redistribute, a structural feature of the field rather than an accident of any one resource. Existing artist-level resources are either tied to proprietary usage data [34] or not curated for music [38]. This leaves open whether artist biographies, deliberately constructed to be high-coverage, reliable, and openly distributable, can occupy the underexplored high-density, high-scalability region. This motivates the first research question and the corpora contributed in Chapter 4.

The second gap concerns the *control* of comparisons. Multimodal MIR demonstrates that adding text helps, but reported gains are confounded by differences in architecture, datasets, and protocol [50], and the field rarely isolates the effect of which text is used under a common pipeline. Whether the benefit of multimodality comes from adding any text at all, or specifically from richer contextual text, cannot be answered without holding the downstream model, fusion strategy, and splits fixed and varying only the representation. Addressing this requires the controlled, representation-centric evaluation framework of Chapter 3 and the systematic comparison of audio, metadata, lyric, and biography representations in Chapter 5.

The third gap concerns the *stage* at which text is used. Prior work leverages text either as an inference-time modality [35, 36, 13] or as a pretraining signal [32, 17, 23, 7], but rarely asks whether knowledge injected explicitly at inference can instead be internalised during training so that its benefit persists when text is unavailable downstream. Because many settings must operate on audio alone, this distinction is of practical as well as scientific importance. Whether artist-level knowledge can be internalised by pretraining an audio-to-text model to generate biographies, and thereby improve audio-only fine-tuning, is the subject of Chapter 6.

In summary, the literature establishes that textual signals can align music representations with higher-level semantics, but leaves unresolved which source can be used at scale and shared openly, how to compare sources under controlled conditions, and whether their benefit can be internalised rather than supplied at inference. These three gaps, of source, control, and stage, structure the remainder of this thesis: Chapter 3 builds the controlled evaluation framework, Chapter 4 constructs and validates the biography corpora, Chapter 5 compares representations to isolate the effect of textual source, and Chapter 6 tests whether artist knowledge can be internalised through generative pretraining.

Chapter 3

Evaluating music representations

Music representation learning lies at the core of modern Music Information Retrieval (MIR) systems. Representations determine which musical properties can be captured, how efficiently models can be trained, and ultimately, which downstream tasks can be solved. As a result, evaluating music representations is a critical methodological challenge: conclusions about a representation's quality are inseparable from the tasks, datasets, and metrics used to assess it.

This chapter introduces the evaluation framework that underpins the empirical contributions of this thesis. The central methodological claim of this chapter is that meaningful comparison of music representations requires a controlled protocol in which the downstream architecture, training procedure, and evaluation settings are held constant, so that observed differences can be attributed primarily to the choice of representation. The evaluation strategy is designed to probe increasingly abstract musical concepts, ranging from categorical descriptors such as genre, to continuous affective dimensions such as valence and arousal, and finally to heterogeneous semantic annotations in music tagging.

Chapter 2 traced the field from hand-crafted descriptors through supervised and self-supervised audio encoders to text-augmented multimodal systems, and identified two obstacles that this chapter responds to directly. The first is that comparisons across the literature are confounded by differences in model architecture, datasets, and evaluation protocol, so a reported gain cannot be attributed to the representation that supposedly produced it. The second is that the benchmarks themselves vary in availability and in the confounds they carry, which limits what any single result can establish. This chapter responds by fixing the architecture, datasets, splits, and metrics, so that the representation becomes the only variable. It begins by formalising the notion of a music representation, distinguishing between raw multimodal inputs and derived representations used by downstream models. It then motivates the use of specific semantic tasks as proxies for musical understanding and discusses the impli-

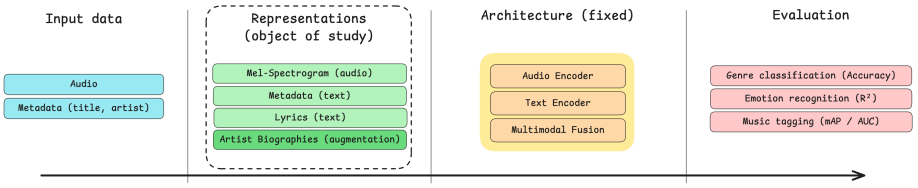


Figure 3.1: **Overview of the evaluation pipeline used throughout this thesis.** Raw audio and textual metadata are transformed into music representations, which constitute the primary object of study. All representations are evaluated using a fixed model architecture across three downstream semantic tasks: genre classification, emotion recognition, and music tagging. Each is associated with standard datasets and evaluation metrics. This design controls for the effect of model architecture, so that differences in performance can be attributed primarily to the choice of representation.

cations of fixing the model architecture in order to control for architectural variation and focus evaluation on the representation itself.

Three evaluation tasks are considered: genre classification, emotion recognition, and music tagging. For each task, the chapter details the corresponding datasets, annotation characteristics, and evaluation metrics, with particular attention paid to dataset balance, annotation quality, and comparability with prior work. Together, these tasks define a unified benchmark for evaluating whether a representation captures transferable semantic information across multiple MIR problem formulations. This evaluation pipeline is used consistently throughout the subsequent experimental chapters. By fixing the tasks, datasets, metrics, and model architecture, the chapter enables a controlled comparison of the representation strategies introduced in Chapter 5.

Figure 3.1 summarises the evaluation pipeline defined in this chapter. Concretely, the framework comprises four components:

- **Input data**: raw audio waveforms and associated textual metadata (e.g., title, artist name), which constitute the available information for each music item;
- **Representations** (object of study): transformations that convert raw inputs into concrete signals suitable for model consumption, such as mel-spectrograms from audio, or templated textual descriptions from metadata and artist biographies (Section 3.1). These are not learned embeddings, but rather the structured inputs presented to the model.
- **Architecture** (fixed): a shared model comprising modality-specific encoders and a multimodal fusion module, held constant across all experiments so that performance differences reflect the choice of representation rather than model capacity;
- **Evaluation**: genre classification (accuracy), emotion recognition (R^2), and music tagging (mAP/AUC), spanning six benchmark datasets of varying scale, annotation type, and artist population.

Dataset	Task	Tracks	Labels	Metric
GTZAN	Genre	1,000	10 genres (single-label)	Accuracy
FMA-Small	Genre	8,000	8 genres (single-label)	Accuracy
EmoMusic	Emotion	744	valence, arousal (continuous)	Avg. R^2
Deezer	Emotion	18,644	valence, arousal (continuous)	Avg. R^2
MagnaTagATune	Tagging	25,860	top 50 tags (multi-label)	mAP / ROC-AUC
Jamendo-MoodTheme	Tagging	17,982	56 mood/theme tags (multi-label)	mAP / ROC-AUC

Table 3.1: **Summary of the six benchmark datasets used in the evaluation framework.** The datasets span three tasks, a wide range of scales, and heterogeneous annotation regimes. Annotation sources range from expert-curated (GTZAN, FMA, Jamendo) to crowdsourced (EmoMusic, MagnaTagATune) and automatically derived from tags (Deezer). Public audio is available for all datasets except Deezer, whose experiments rely on proprietary audio.

This framework provides the methodological basis for the empirical studies presented in the remainder of the thesis. By fixing the downstream conditions of evaluation, the subsequent chapters can attribute performance differences primarily to the representations themselves, enabling controlled comparison of audio, textual, and multimodal approaches.

Table 3.1 summarises the six benchmark datasets that instantiate this evaluation across the three tasks; each is described in detail in the corresponding task section below. These are the same datasets surveyed in Chapter 2, where Table 2.1 assessed them as *resources*, in terms of availability and the confounds they carry. Here they are set out as an *evaluation protocol*: what each one asks a model to predict, over which label space, and under which metric. The set spans a range of scales (from 1,000 to over 25,000 tracks), annotation regimes (expert-curated, crowdsourced, and automatically derived labels), and label types (single-label genres, continuous affective dimensions, and multi-label tags), so that conclusions about a representation do not rest on the idiosyncrasies of any single benchmark.

3.1 Music representation framework

In this section, we define what is meant by a *music representation*, distinguishing between *raw inputs*, *derived representations*, and their use by *downstream predictors*. We begin by describing the raw input space. A music item is represented by an audio waveform together with associated textual metadata.

Formally, let $w \in \mathbb{R}^{s_r \times \tau}$ denote the waveform of a song, where s_r is the sampling rate and τ its duration in seconds. Let $x_t \in V^n$ denote the track title as a sequence of n tokens drawn from a vocabulary V, and let $x_a \in V^m$ denote the artist name as a sequence of m tokens from the same vocabulary. The minimal multimodal input considered in this thesis is therefore the tuple (w, x_t, x_a). This formulation provides the basis for defining representation functions that transform raw multimodal inputs into features suitable for downstream prediction tasks.

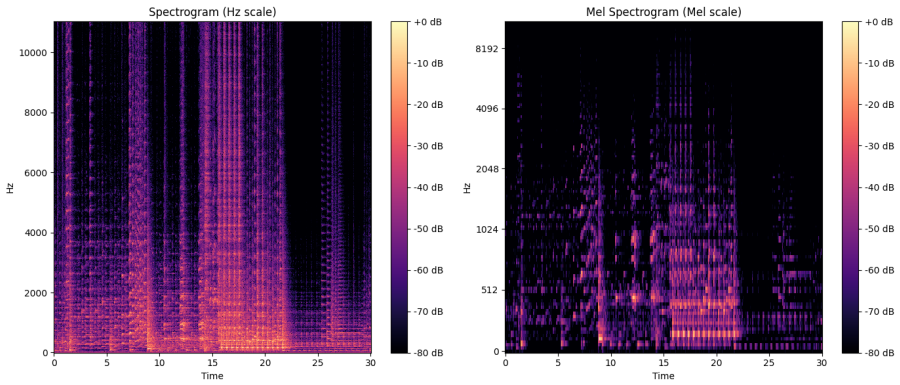


Figure 3.2: Example waveform-derived time-frequency representations of `jazz.00000.wav` from the GTZAN dataset [49], generated using the Librosa library. The Mel spectrogram provides a compact, perceptually motivated representation commonly used in MIR systems.

A music representation is then defined as a transformation function:

$$g : \mathbb{R}^{s_r \times \tau} \times V^n \times V^m \to R_m, \tag{3.1}$$

where R_m denotes the multimodal representation space. This definition is intentionally general: depending on the representation method, R_m may consist of an audio representation alone, a textual representation alone, or a joint multimodal representation combining both.

In this thesis, a music representation is understood as the set of information provided to a model for a given music item, before any model-specific encoding. This includes the raw audio signal, associated metadata, and any additional information used to augment the input. The goal of such representations is to capture both low-level acoustic structure and higher-level properties that can be associated with the music item, in order to support downstream semantic tasks.

This formulation makes the representation function itself the primary object of evaluation, which is essential for comparing audio-only, textual, and multimodal inputs within a common framework.

In practice, MIR systems rarely operate directly on the raw waveform. Instead, they typically rely on derived audio representations that transform the waveform into a more compact and structured form. A common example is the Mel spectrogram [9], a time-frequency representation obtained by applying a short-time Fourier transform over successive temporal windows and projecting the frequency axis onto the Mel scale; Figure 3.2 shows an example for a track from GTZAN. We adopt the Mel spectrogram as the default audio representation for several reasons. First, the Mel scale is logarithmic in frequency and therefore approximates the frequency resolution of human hearing, allocating

more resolution to the lower frequencies where pitch and timbral information are perceptually concentrated. Second, projecting the linear frequency axis onto a small number of Mel bands substantially reduces dimensionality relative to the raw STFT while preserving the harmonic and timbral structure most relevant to the semantic tasks considered here. Third, the Mel spectrogram is the de facto standard input to convolutional audio models in MIR, as surveyed in Chapter 2; adopting it therefore keeps our audio representation comparable with prior work rather than introducing an idiosyncratic front-end. More generally, any derived audio representation can be written as a transformation:

$$g_w : \mathbb{R}^{s_r \times \tau} \to R_w, \tag{3.2}$$

where R_w is the audio representation space. For Mel spectrograms, for example, $R_w = \mathbb{R}^{b \times t}$, where b is the number of Mel bins and t the number of temporal frames.

A downstream model then maps the chosen representation to a task-specific output. Let

$$f : R_m \to \mathbb{R}^o, \tag{3.3}$$

where o denotes the output dimensionality required by the task. The full prediction pipeline can therefore be expressed as the composition:

$$f \circ g : \mathbb{R}^{s_r \times \tau} \times V^n \times V^m \to R_m \to \mathbb{R}^o. \tag{3.4}$$

This distinction is central to the methodology of this thesis. The object of study is the representation function g, while the downstream predictor f is held fixed across experiments. This allows the evaluation to focus on how different representation choices affect downstream performance, rather than conflating representation quality with architectural variation. In Chapter 5, this framework is instantiated using the specific audio, textual, and multimodal representations that are compared throughout the thesis.

3.1.1 What are we evaluating?

In music information retrieval, the quality of a representation is typically assessed through its performance on downstream tasks. Because many musical properties of interest are semantic and context-dependent, no single task can fully characterise representation quality. Instead, this thesis evaluates representations across several complementary semantic prediction problems.

We consider three tasks: *genre classification*, *emotion recognition*, and *music tagging*. Genre classification evaluates whether a representation captures stylistic categories associated with musical identity. Emotion recognition probes affective interpretation by mapping songs onto continuous valence–arousal dimensions. Music tagging provides a broader multi-label setting in which tracks are associated with heterogeneous descriptors, including genres, moods, instrumentation, and other semantic attributes.

Together, these tasks probe different aspects of musical meaning. Genre classification focuses on categorical stylistic identity, emotion recognition captures subjective affective responses, and music tagging evaluates the ability to model heterogeneous semantic annotations under strong label imbalance. This combination provides a diverse yet complementary evaluation suite for testing music representations.

The chosen tasks also span different prediction settings: single-label classification (genre), regression (emotion), and multi-label classification (tagging). This diversity requires the evaluation pipeline to support multiple task formulations while maintaining a consistent representation interface.

Finally, all datasets used in this chapter include metadata fields such as track title and artist name. These attributes are required for the multimodal representations explored in later chapters and ensure that the evaluation pipeline remains compatible with both audio-only and multimodal representation learning approaches.

3.1.2 Music representations vs model architecture

The framework introduced in Section 3.1 separates the representation function g from the downstream predictor f. In practice, a prediction is obtained by first transforming the raw multimodal inputs into a representation and then applying a task-specific model. This pipeline can therefore be expressed as the composition:

$$f \circ g : \mathbb{R}^{s_r \times \tau} \times V^n \times V^m \to R_m \to \mathbb{R}^o. \tag{3.5}$$

This formulation exposes a central methodological problem in MIR evaluation: representation quality cannot be meaningfully compared when the downstream model changes simultaneously. In such cases, improvements may reflect architectural advantages rather than the semantic quality of the representation. To address this, the evaluation framework proposed in this chapter fixes the downstream predictor across all experiments. It treats the representation function g as the primary object of study, while f serves as a controlled predictor that maps representations to task outputs.

For this purpose, we adopt a fixed multimodal architecture that is held constant across experiments. The architecture comprises modality-specific encoders, a shared fusion module, and task-specific prediction heads, so that the primary source of variation lies in the representation rather than the downstream model. In the instantiation used in Chapter 5, VGGish [21] serves as the audio encoder. The choice of VGGish follows directly from the methodological aim of this framework. Because the encoder is deliberately *not* the object of study, the priority is a stable, well-understood, and widely adopted audio encoder rather than the most recent state-of-the-art model: a moving encoder would reintroduce exactly the architectural confound the framework is designed to remove. VGGish satisfies this requirement well. It is pretrained on the large-scale AudioSet corpus, producing compact, general-purpose audio embeddings that have been used extensively as a baseline audio front-end in MIR

(Chapter 2), which makes our results comparable with prior work and cheap to reproduce. Holding this encoder fixed across all tasks and datasets ensures that differences in downstream performance can be attributed primarily to the representation being evaluated rather than to encoder capacity. Figure 3.1 illustrates the resulting evaluation pipeline, highlighting the separation between the fixed downstream architecture and the variable representation component.

3.2 Genre classification

This task is a single-label classification problem in which a model classifies a song into a genre label. This task assumes a known, finite set of labels. The datasets we selected are balanced across all classes.

Genre classification evaluates whether a representation captures high-level stylistic structure associated with musical identity. Although it is a standard MIR benchmark, it is particularly relevant here because genre depends on multiple factors, including instrumentation, era, style, and the artist's cultural background. It therefore provides a test of semantics that are partly acoustic and partly culturally mediated.

3.2.1 Evaluation

We evaluate model performance using classification accuracy, defined as the proportion of correctly classified samples over the total number of samples in the dataset. Let $y_i \in Y$ denote the ground-truth label for sample i, and let $\hat{y}_i$ be the corresponding predicted label. We define accuracy as:

$$Accuracy = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i), \tag{3.6}$$

where N is the total number of samples and $\mathbb{I}(\cdot)$ is the indicator function.

Accuracy is intuitive because it corresponds directly to the question one would naturally ask of a genre classifier, "what fraction of songs did it label correctly?", and reports it on a single 0–1 scale that requires no threshold tuning, class weighting, or familiarity with information-retrieval metrics to interpret. Equivalently, it estimates the probability that a randomly chosen song is assigned its correct genre. This interpretability is, however, only trustworthy when the classes are balanced. Under class imbalance, a model can achieve high accuracy simply by favouring the majority class while performing poorly on underrepresented genres, so the headline number no longer reflects performance across the full label space. In other words, accuracy is intuitive by construction but only *valid* as a summary of performance when no class dominates. We therefore restrict accuracy-based evaluation to datasets with balanced genre distributions, where its intuitive reading and its validity coincide.

To mitigate these limitations and ensure fair comparison with prior work, we rely on datasets with balanced genre labels and use the publicly available

training and test splits provided by the dataset creators. These splits ensure consistency in the evaluation protocol and comparability with previously reported results.

3.2.2 Datasets

GTZAN

The GTZAN dataset is a well-established collection that has been extensively analysed and re-annotated over the years. It is made of $1,000$ songs across the following genres:

Blues	Country	Hip-Hop	Metal	Reggae
Classical	Disco	Jazz	Pop	Rock

Each genre has 100 samples, and each sample is a $22,050$ Hz, mono-channel, 16-bit audio file. The dataset does not include any metadata (title, artist, album), but Sturm [47] provides approximately 90% of them.

A notable characteristic of this dataset is the impact of the choice of train-dev-test split. The initially provided train-test split was random. The main information "leak" was that most artists appeared in both the training and test splits, providing the model with a few examples for each. To create a more challenging scenario, [26] provides an "adversarial split" in which each artist appears only in either the training or the testing set. We use this split for our evaluation to test our model in a more challenging environment.

Another element worth highlighting in this dataset is the choice of genres. The primary driver of these choices was the coverage of music genres in 2002. However, some of these genres present a bigger challenge than others. In this setup, most models do not struggle to classify classical music, but there are many more nuances in hip-hop, pop, and rock.

This dataset was originally hosted here[1], but its hosting has been discontinued. The audio data can now be found here[2]. As part of the evaluation pipeline, we have released the metadata and split used[3].

Free Music Archive - Small

The Free Music Archive is a collection of 106,574 tracks across 161 unbalanced genres. From this dataset, three subsets are derived.

- **Small**: $8,000$ samples across eight balanced genres
- **Medium**: $25,000$ samples across 16 unbalanced genres
- **Large**: $106,574$ samples across 161 unbalanced genres

Given our need for balanced genres, we choose to use the Small subset. The genres are:

[1] `http://marsyas.info/downloads/datasets.html`

[2] `https://huggingface.co/datasets/yangwang825/gtzan`

[3] `https://github.com/fedingo/music_multimodal_augmentations/blob/main/tasks/data/GTZAN_official.json`

Hip-Hop	Folk	Rock	Electronic
Pop	Experimental	International	Instrumental

Each of the 8,000 samples is provided at a sampling rate of 44,100 Hz and in stereo. The dataset includes detailed metadata for each song, such as the title, artist, album, and other attributes. This information will be required for the development of our textual representations described in Chapter 5.

This dataset presents a significantly greater challenge than GTZAN due to its being eight times larger and having much closer genre overlap.

The audio data is available for all subsets at this link[4] We release, as part of our evaluation pipeline, the metadata and splits we have used.[5]

3.3 Emotion recognition

Emotion recognition is a regression problem that maps each song onto two axes: valence and arousal. Valence describes whether an emotion is pleasant (positive) or unpleasant (negative), while arousal measures its intensity, from calm (low arousal) to excited (high arousal). These two axes are the coordinates of Russell's *circumplex model of affect* [44], shown in Figure 3.3, which we adopt as the emotional representation throughout this thesis.

We adopt the circumplex model for two reasons. First, and most importantly, it is the de facto standard for representing emotion in affective MIR: the datasets used here are annotated in valence–arousal space, and reporting results along these axes keeps our evaluation consistent with their ground-truth annotations and comparable with the body of prior work reviewed in Chapter 2. Second, its continuous formulation deliberately introduces a non-categorical task into the benchmark. Whereas genre classification and tagging are categorical, the circumplex casts emotion as a regression problem, allowing us to probe whether a representation captures graded, continuous affective information rather than only discrete class membership. This broadens the range of prediction settings against which representations are tested.

This problem is relevant because it evaluates the ability of music representations to capture the mood and emotions evoked by a song. One crucial challenge with this task is its high subjectivity.

Emotion recognition provides a complementary test of representation quality because it probes affective semantics, which are more subjective and often less directly inferable from acoustics alone.

3.3.1 Evaluation

Emotion recognition typically uses the coefficient of determination (R^2) to evaluate model performance. It measures how well predictions reproduce ground-

[4] `https://github.com/mdeff/fma`

[5] `https://github.com/fedingo/music_multimodal_augmentations/blob/main/tasks/data/fma_genre_filtered.json`

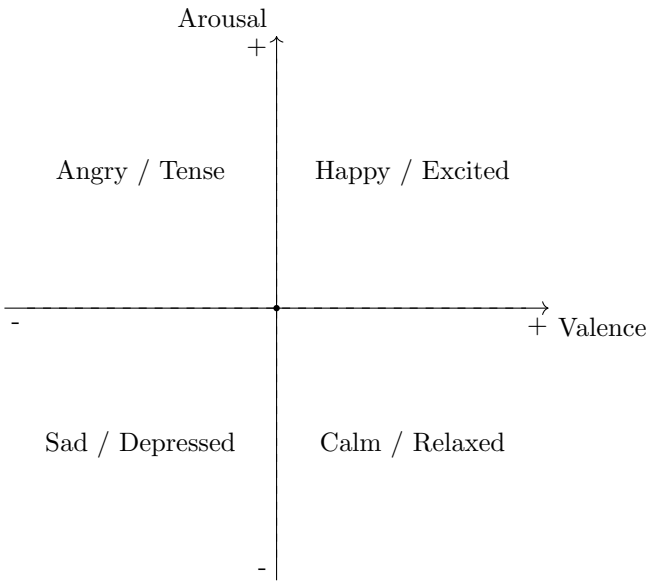


Figure 3.3: Valence–arousal space commonly used to represent music emotion [44].

truth labels by using the proportion of variation explained by the predictions. Let y_i denote the ground-truth value for sample i, and let $f_i \in f(X)$ denote the corresponding model prediction. The coefficient of determination is defined as:

$$R^2 = 1 - \frac{\sum_i (f_i - y_i)^2}{\sum_i (y_i - \overline{y})^2}. \tag{3.7}$$

This can be interpreted as 1 minus the ratio of the average residual (model errors) to the data variance. Using this score, we will obtain a value for each predicted field (valence and arousal). To obtain a single comparable value, we take the arithmetic mean of the two per-dimension scores, weighting valence and arousal equally, since neither axis is considered more important than the other:

$$Avg.R^2 = \frac{R^2_{valence} + R^2_{arousal}}{2}. \tag{3.8}$$

3.3.2 Datasets

EmoMusic

This dataset was created in 2013 as part of the MediaEval competition. The audio tracks were selected from the Free Music Archive through a subset selection from Soleymani et al. [45]. They selected the top 300 songs from the following genres: blues, electronic, rock, classical, folk, jazz, country, and pop. They then

filtered the collection by audio file length (more than 1 minute or less than 10 minutes) and selected the top 125 songs for each genre.

Using this 1,000-song collection, they run a crowdsourced campaign collecting annotations from 100 selected Mechanical Turk workers. The annotations were collected as continuous annotations, in which the worker moved a slider while listening to the song and produced annotations over time. In the final dataset, redundancy was detected, reducing the corpus to 744 songs. The authors provide a train-test split comprising 619 and 125 samples, respectively.

The authors report a high level of noise caused by 2 factors. One was technological, with their annotation interface sometimes failing to capture the workers' slider movements. The second was due to workers' lack of understanding or engagement during the annotation process.

For our task, we convert the continuous, time-varying annotations into a single static target per song: we take the arithmetic mean of the valence trajectory and of the arousal trajectory over the annotated duration, and predict these two mean values. We use the mean, rather than the median, as it is the standard aggregation in prior static emotion-regression work, which preserves comparability with previously reported results. This transformation reduces the original temporal modelling problem to a static regression task.

The audio data is available upon request at this link[6]. As part of our evaluation pipeline, we release the split and metadata used.[7]

Deezer

This dataset is derived from the Million Song Dataset [4], composed of 18,644 songs. They annotate these songs on the valence/arousal scale using the following automated process. They use a subset of the provided tags for each song that are related to mood, and then use Warriner et al. [51] to map each tag to the valence–arousal score. If more than 1 tag is assigned to a song, they average the score of all the tags. This process allows the creation of a large-scale dataset for emotion recognition, at the expense of lower-quality annotations.

The authors create a train-dev-test split via random sampling, with the only constraint that songs from the same artist will not appear in different sets. The resulting dataset consists of $18,644$ samples, split into $11,267$ for the training set, $3,863$ for the validation set, and $3,514$ for the test set.

This dataset has properties opposite to EmoMusic: we have more than 20 times as much data, but at the expense of annotation quality. This difference is reflected in reported performance levels, with substantially higher R^2 scores on EmoMusic than on Deezer. EmoMusic's best model has 0.669, while Deezer has 0.207. These numbers are not directly comparable, but they do indicate a difference in the quality of the annotations provided.

The audio data for this dataset is not publicly available. The authors of

[6] https://cvml.unige.ch/databases/emoMusic/

[7] https://github.com/fedingo/music_multimodal_augmentations/blob/main/tasks/data/emoMusic.json

this dataset have released the splits used.[8] However, it relies on having access to the Million Song Dataset. This dataset has been publicly released `http://millionsongdataset.com/`. However, this release uses an encoded version of the audio files in a feature format called EchoNest, which encodes the audio into "pitch segments" and "timbre segments". For our experiments, we have used a proprietary dataset provided by the company that sponsored this work (MoodAgent). As part of the pipeline, we release the metadata and splits we used[9].

3.4 Music tagging

Music Tagging is a multi-label classification problem that assigns a set of tags to each song. Most datasets have a broad range of possible tags, spanning emotional categories (e.g., happy or energetic), genres (e.g., electronic or pop), and instruments (e.g., violin or guitar). This problem blends emotion recognition and genre classification in a very complex setup.

Music tagging completes the benchmark by testing whether a representation supports heterogeneous and highly imbalanced semantic annotations, thereby approximating the broader descriptive vocabulary used in real-world MIR settings.

3.4.1 Evaluation

To evaluate music tagging performance, we use two metrics: mean average precision (mAP) and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). While all models are optimised using mAP during validation, we report both mAP and ROC-AUC for completeness and comparability with prior work.

Mean Average Precision

Mean average precision (mAP) is a commonly used evaluation metric for multi-label classification tasks, particularly in scenarios with strong class imbalance, such as music tagging. Unlike accuracy-based metrics, mAP evaluates the quality of the predicted confidence scores across all possible decision thresholds.

Let $y_{i,t} \in \{0, 1\}$ denote the ground-truth label indicating the presence or absence of tag t for sample i, and let $s_{i,t} \in \mathbb{R}$ be the corresponding confidence score predicted by the model. For each tag t, a precision–recall (PR) curve is obtained by varying the decision threshold applied to $s_{i,t}$.

The average precision (AP) for tag t is defined as the Area under the corresponding precision–recall curve:

[8] `https://github.com/deezer/deezer_mood_detection_dataset`

[9] `https://github.com/fedingo/music_multimodal_augmentations/blob/main/tasks/data/deezer.json`

$$AP_t = \int_0^1 p_t(r)\, dr, \tag{3.9}$$

where $p_t(r)$ denotes the precision as a function of recall for tag t. In practice, this integral is approximated by a finite sum over the recall values at which precision changes:

$$AP_t = \sum_k (R_t(k) - R_t(k-1))\, P_t(k), \tag{3.10}$$

where $P_t(k)$ and $R_t(k)$ denote the precision and recall at the k-th threshold.

The mean average precision is then computed by averaging the AP values across all tags:

$$mAP = \frac{1}{T} \sum_{t=1}^{T} AP_t, \tag{3.11}$$

where T is the total number of tags. This formulation treats each tag independently and provides a threshold-independent measure of classification performance that is robust to label imbalance.

Receiver Operating Characteristic – Area Under the Curve

The Receiver Operating Characteristic (ROC) curve evaluates the trade-off between true positive rate (TPR) and false positive rate (FPR) across all possible classification thresholds. The Area Under the ROC Curve (ROC-AUC) provides a threshold-independent measure of a model's discriminative ability.

For a given tag t, the TPR and FPR are defined as:

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}, \tag{3.12}$$

where TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives, respectively. The ROC curve is obtained by varying the decision threshold applied to $s_{i,t}$.

Similar to the average precision, ROC-AUC for a single tag is computed as:

$$\textit{ROC-AUC}_t = \int_0^1 TPR(FPR) dFPR. \tag{3.13}$$

This definition is equivalent to computing the probability that if we choose two samples $(s_{i,t}, s_{j,t})$ where $y_i = 1$ and $y_j = 0$, then $s_{i,t} > s_{j,t}$. In other words, the $\textit{ROC-AUC}_t$ score for a tag t is equivalent to the probability that the score of a random positive sample is higher than the score of a random negative sample.

The overall ROC-AUC is then obtained by computing the average across all tags:

$$\textit{ROC-AUC} = \frac{1}{T} \sum_{t=1}^{T} \textit{ROC-AUC}_t. \tag{3.14}$$

3.4.2 Datasets

MagnaTagATune

The MagnaTagATune dataset is a widely used benchmark for music tagging and annotation tasks. It was originally introduced as part of an online game designed to collect semantic annotations for music through crowdsourcing. The dataset consists of approximately 25,860 audio clips, each 29 seconds long, sampled at 44,100 kHz and in mono.

Each audio clip is annotated with a set of tags drawn from a vocabulary of 188 tags, covering a broad range of musical attributes, including genre, instrumentation, vocal presence, and mood-related descriptors. The annotations were collected from multiple users, and each tag assignment reflects consensus among annotators rather than individual judgments.

Due to the heterogeneous nature of the tag vocabulary, prior work restricts evaluation to a subset of frequently occurring tags. Following standard practice, we focus on the top 50 tags to reduce sparsity and improve evaluation reliability. As MagnaTagATune is a multi-label dataset with highly imbalanced tag distributions, the chosen evaluation metrics, mean average precision and ROC-AUC, are best suited to these problems.

We use the train-dev-test splits used by Law et al. [25]. This split keeps each sample without tags rather than filtering them out, and removes only three samples with corrupted audio data from the original release. The audio data can be found here [10]. As part of our evaluation pipeline, we release our metadata and splits used[11].

Jamendo-MoodTheme

The Jamendo dataset is a large-scale music tagging corpus derived from the Jamendo music platform and released as part of the MediaEval benchmarking initiative. It comprises 56,639 full-length tracks annotated with semantic tags covering genres, instruments, and mood-related concepts. The dataset is notable for its scale and diversity, as well as for the use of professionally curated annotations.

From this collection, we use the Jamendo-MoodTheme subset, which focuses specifically on mood and theme annotations. This subset contains 17,982 tracks annotated with a reduced vocabulary of 56 mood- and theme-related tags. The annotations are provided in a multi-label format, where each track can be associated with multiple mood or theme descriptors.

Compared to MagnaTagATune, Jamendo-MoodTheme uses a more abstract, semantically overlapping vocabulary, which increases task difficulty. Additionally, while annotations are curated, they are not derived from continuous affective judgments, but rather from categorical mood descriptors.

[10] `https://mirg.city.ac.uk/codeapps/the-magnatagatune-dataset`

[11] `https://github.com/fedingo/music_multimodal_augmentations/blob/main/tasks/data/mtat_unfiltered.json`

We use the `split-0` training-test split to maintain consistency with prior evaluations. As with MagnaTagATune, we treat music tagging as a multi-label classification problem and report performance using mean average precision and ROC-AUC.

The audio data can be found here.[12] As part of our evaluation pipeline, we release our metadata and splits used.[13]

3.5 Contributions

This chapter makes a methodological contribution to music representation learning by defining a controlled and reusable evaluation framework for comparing representations across semantic MIR tasks. Rather than varying model architectures across datasets, the proposed framework holds the downstream predictor, training protocol, and evaluation settings fixed, allowing performance differences to be attributed primarily to the choice of representation.

The implementation of the evaluation pipeline is publicly available[14], including the dataset splits and metadata used in our experiments[15]. Due to copyright restrictions on some datasets, only the experiments on GTZAN, FMA-Small, EmoMusic, MagnaTagATune, and Jamendo-MoodTheme can be fully reproduced without access to proprietary audio data.

The main contributions of this chapter are:

- **A controlled evaluation framework for music representation learning.** We formalise music representation evaluation as the comparison of representation functions under a fixed downstream predictor and shared training protocol. This design controls for architectural variation, so that observed differences in performance can be attributed primarily to the choice of representation.
- **A unified cross-task benchmark for semantic MIR evaluation.** We define a benchmark spanning genre classification, emotion recognition, and music tagging, covering single-label classification, regression, and multi-label classification settings. This enables representation quality to be assessed across complementary forms of musical semantics rather than on a single task alone.
- **A reproducible experimental protocol and released benchmark resources.** We consolidate six widely used MIR datasets into a single evaluation pipeline and release the metadata, splits, and implementation required to reproduce the benchmark wherever licensing permits.

[12] `https://mtg.github.io/mtg-jamendo-dataset/#downloading-the-data`

[13] `https://github.com/fedingo/music_multimodal_augmentations/blob/main/tasks/data/jamendo-mt_low_crop.json`

[14] `https://github.com/fedingo/music_multimodal_augmentations`

[15] `https://github.com/fedingo/music_multimodal_augmentations/tree/main/tasks/data`

3.6 Summary

This chapter defines the evaluation framework used throughout this thesis to assess the quality of music representations. It formalises music representations as transformations from raw multimodal inputs (audio waveforms and associated textual metadata) into representation spaces that support downstream prediction tasks. The chapter also clarifies the distinction between representation learning and model architecture, emphasising that the downstream predictor is held fixed in order to control for architectural variation and focus evaluation on the representation itself.

To evaluate representation quality, the chapter introduces a task-driven evaluation pipeline based on three complementary MIR problems: genre classification, emotion recognition, and music tagging. These tasks probe different aspects of musical semantics, ranging from categorical stylistic labels to affective interpretation and heterogeneous multi-label annotations. Together, they provide a diverse yet coherent benchmark for assessing music representations.

The resulting framework consolidates multiple widely used MIR datasets and evaluation metrics into a single controlled experimental pipeline. By defining music representation evaluation as a controlled comparison under fixed downstream conditions, this chapter provides the methodological basis for the thesis's later empirical claims. It positions the benchmark itself as a contribution: a reusable protocol for assessing whether different information sources improve semantic music representations across tasks.

Chapter 4

Artist biography corpora

Chapter 3 established the evaluation framework used throughout this thesis, enabling controlled comparison of music representations across multiple downstream music information retrieval (MIR) tasks. Building on that framework, this chapter introduces *augmented representations*, in which the raw music inputs are complemented with external knowledge sources. In this thesis, external knowledge refers to information that is not contained in the original music item itself, but can provide additional task-agnostic semantic context.

In MIR, textual information has been shown to complement audio for tasks such as genre classification, emotion recognition, and music tagging, particularly when supervision is weak or labels are noisy. In practice, however, the use of external text is constrained by two recurring challenges: limited coverage across datasets and uncertainty about the reliability of the information being introduced.

This chapter addresses these challenges by focusing on artist biographies as a source of external knowledge for music representation learning. Artist biographies offer a practical trade-off between coverage and information provided. Compared with song-level textual resources, which are often sparse and constrained by copyright, artist-level descriptions are more widely available and can encode stylistic, geographical, and historical contexts relevant to multiple MIR tasks.

The goal of this chapter is therefore twofold: first, to construct reproducible corpora of artist biographies suitable for representation augmentation; and second, to characterise their properties and limitations. To this end, we introduce two complementary biography corpora: one derived from web-scraped, community-authored sources and one generated using a large language model to produce concise, stylistically homogeneous descriptions. We then evaluate their factual reliability through an extractive question-answering protocol grounded in Wikidata, allowing us to compare how community-authored and LLM-generated biographies encode verifiable artist information.

Together, these analyses establish the resources used in the remainder of the thesis and clarify the trade-offs between coverage, homogeneity, and factual

reliability that must be considered when augmenting music representations with artist-level text.

4.1 Augmenting music representations with external knowledge

In Section 3.1, we define a music representation as a function g that maps (w, x_t, x_a) to a representation space R_m. There, we clarify that a music representation corresponds to the set of information provided to a model for a given music item. This formulation assumes that all information used to construct the representation is contained within the initial input triplet, without access to any other sources.

We extend this definition by allowing representations to incorporate information retrieved from external sources during their construction. In this view, augmentation can be understood as expanding the set of information available to the model beyond what is contained in the original music item.

We define an **augmented music representation** as an extension of g that incorporates external knowledge retrieved from an auxiliary source. We model this augmentation through a function k that maps the available inputs to an associated textual representation:

$$k : \mathbb{R}^{s_r \times \tau} \times V^n \times V^m \to V^d. \tag{4.1}$$

Depending on the source, k may correspond to:

- Retrieval from an external corpus (e.g., artist biographies)
- Transformation of the input signal (e.g., speech-to-text transcription)

The augmented representation function $\hat{g}$ can then be defined as an extension of g that incorporates the additional textual input produced by k:

$$\hat{g} : \mathbb{R}^{s_r \times \tau} \times V^n \times V^m \times V^d \to R_m^a, \tag{4.2}$$

where R_m^a represents the augmented multimodal representation space.

An important factor in retrieval-based augmentation is coverage. In our case, we define **coverage** as the proportion of samples for which the source provides non-empty information:

$$coverage(K) = \frac{1}{N} \sum_i \mathbb{I}[k(w^i, x_t^i, x_a^i) \neq \emptyset]. \tag{4.3}$$

In practice, coverage reflects whether a knowledge source contributes additional information to the representation. For example, missing lyrics correspond to zero coverage, whereas available text can enrich the representation. This definition, therefore, provides a unified way to compare heterogeneous textual sources.

In this thesis, we consider three different types of external knowledge sources:

- **Curated encyclopedic knowledge**: These sources provide high-quality information, but suffer from coverage gaps and copyright constraints.
- **Community-created annotations**: These sources provide a much wider coverage, but suffer from quality issues and stylistic heterogeneity.
- **LLM-generated knowledge**: These sources rely on the knowledge embedded within an LLM. They have a standardised style but run the risk of hallucinating information, which can lead to augmentation with false information.

4.1.1 Why artist biographies

For augmentation to have a measurable impact, the underlying knowledge source must provide sufficient coverage across the evaluation datasets. Song-level textual resources are often sparse, unevenly distributed, and frequently constrained by copyright (e.g., lyrics).

Artist-level descriptions provide a more stable and widely available source of contextual information, capturing stylistic, geographical, and historical attributes that are relevant across multiple tracks.

Existing artist-level resources, such as Wikipedia pages or metadata collections derived from the Million Song Dataset (e.g., MSD-Artist [34]), provide useful contextual information, but each falls short of our requirements in specific ways. Wikipedia's notability criteria mean that the independent and long-tail artists who dominate research datasets such as the Free Music Archive and Jamendo are simply absent: when Wikipedia passages are retrieved for the artists in our evaluation datasets, only 41.0% of samples obtain a usable biography overall, dropping to 7.2% on Jamendo-MoodTheme (we quantify this in Chapter 5, Table 5.1). MSD-Artist inherits a related bias: it is anchored to the mainstream-oriented Million Song Dataset, covers 36.5% of our evaluation samples overall (0.8% on Jamendo-MoodTheme), and its descriptions are short (26 tokens on average), limiting the semantic context they can provide. Both resources are also fixed snapshots that cannot be extended to new artists appearing in a given dataset. To address these issues, we construct two new artist biography corpora designed to provide both high coverage across the artist populations actually present in MIR benchmarks and controlled textual representations.

4.2 Proposed corpora

4.2.1 Web-scraped biographies

Corpus construction

To create this corpus, we collect artist biographies from platforms directly associated with the datasets used in our evaluation pipeline:

- `https://jamendo.com`
- `https://freemusicarchive.org`
- `https://magnatune.com/`

These sources are selected for two reasons. First, they are the original sources of the audio datasets used in this thesis, thereby maximising coverage and ensuring direct alignment. Second, they provide publicly accessible, user-authored artist pages, enabling the construction of a reproducible corpus.

An additional advantage of these sources is that they contain community-authored biographies written for self-promotion. While this introduces variability and noise, it reflects realistic textual descriptions encountered in music platforms and provides a complementary perspective to more structured or curated knowledge sources.

We leverage the exports provided by the services (Tarballs or SQLite DBs) to extract the artists' data. Our initial extraction collects 16,916 biographies from the Free Music Archive [12], 754 from a DB export by MagnaTune[1], and 1,508 from Jamendo [6]. This initial extraction retains all available fields, including multilingual biographies, images, external links, and associated tags.

After extraction, we filter out biographies that are shorter than a full sentence or consist only of non-informative content (e.g., emojis). We then deduplicate entries by artist stage name, yielding a corpus of 18,912 artist entries with an average length of 136.1 tokens and a standard deviation of 408.5 tokens.[2] As Section 4.2.1 details, that average is taken over all entries, a majority of which carry no biography text at all; conditioned on the entries that do carry text, the mean is substantially higher.

The dataset metadata, including artist names, source identifiers, tags, and external links, has been publicly released with the publication [43][3]. The raw biography texts are not redistributed, as the underlying platforms provide content under a mix of Creative Commons and platform-specific licences that do not uniformly permit redistribution of textual content. The corpus can be reproduced by parsing the FMA metadata from [12], the Magnatune SQLite export[4], and the Jamendo API[5]. The extracted data can then be mapped back to the provided corpus, and the statistics reported in this section are computed over the collection recovered in this way.

Initial analysis

A key design choice in constructing this corpus is using the artist's stage name as the primary identifier for linking biographies to audio samples. This choice is motivated by the need to maximise overlap across datasets, as each source (e.g., Million Song Dataset, Jamendo, Free Music Archive) uses different internal

[1] `http://magnatune.com/info/sqlite-normalized`

[2] All token counts reported for both corpora in this chapter are computed with the `cl100k_base` byte-pair encoding, the tokeniser of the `gpt-3.5-turbo` model used to build the LLM corpus, applied to the raw stored text (including any HTML markup). We state the tokeniser explicitly because subword tokenisers differ by 10–20% on this kind of text, and the figures are therefore not comparable across tokenisers.

[3] `https://github.com/fedingo/music_multimodal_augmentations/blob/main/models/text_representation/data/author_corpus.json`

[4] `http://magnatune.com/info/sqlite-normalized`

[5] `https://api.jamendo.com`

identifiers.

However, this approach introduces two sources of ambiguity: (i) variation in name spelling or formatting (e.g., acronyms, capitalisation), and (ii) the presence of multiple artists sharing the same name.

To assess the first issue, we perform an edit-distance analysis to identify potentially duplicated or closely matching artist names. We observe that for names longer than five characters, no pairs exhibit small edit distances indicative of likely duplication. Short names (fewer than six characters) produce a higher number of false positives (e.g., "MAKO" vs "PAKO"), and are therefore excluded from this analysis. Overall, this suggests that spelling variation is not a major source of duplication in the corpus.

The second issue, name ambiguity across distinct artists, cannot be fully resolved without access to external entity identifiers. In this work, we accept this limitation and treat artist names as dataset-specific identifiers. This assumption is justified by the fact that all datasets used in this thesis are point-in-time snapshots, in which artist names are consistently associated with the corresponding audio. As a result, any ambiguity is likely to be systematic rather than random, meaning that a given artist will either consistently match or consistently fail to match across the pipeline.

This design prioritises coverage and simplicity over strict entity resolution, which is consistent with constructing large-scale, reproducible corpora.

We report in Figure 4.1 examples of the resulting data objects.

As we can see from the 2 samples, the data contains the following fields:

- **author_name**: contains the artist name, used as a unique identifier to link it to our dataset samples.
- **biographies**: contains a list of biographies from the different sources for the artist.
- **images**: contains a list of URLs of images that were part of the data in the original extraction.
- **tags**: contains a set of tags that were attached to the artist.
- **unique_id**: contains a list of string IDs coming from the different sources. 120 samples have 2 distinct IDs, and 4 samples have 3 distinct IDs (from the three different sources).
- **website**: contains a list of URLs to personal websites of the artist.

A property worth discussing in this dataset is the heterogeneous distribution of biographies. As we can see from Figure 4.1, some biographies are very in-depth, while others contain only a few vague words. This is because they come from a community effort, with no specific requirements or guidelines for what the content must include. We quantify this heterogeneity in the following section.

Descriptive statistics

Because this heterogeneity is the defining property of the corpus, summary means describe it poorly. This section characterises the collection in more detail, since its sparsity and skew condition how the results in Chapters 5 and 6 should be read.

```
{
    'author_name': 'AWOL',
    'biographies': ['<p>A Way Of Life, A Collective of Hip-
       Hop from NJ.....</p>'],
    'image': ['https://freemusicarchive.org/file/images/
       artists/AWOL_-_2009113014453413.jpg'],
    'tags': ['awol'],
    'unique_id': ['1'],
    'website': ['http://www.AzillionRecords.blogspot.com']
}

{'author_name': 'Chandeliers',
 'biographies': ['<p><span style="font-family:Verdana,
     Geneva, Arial;font-size:xx-small;">Chandeliers drink the
      same Chicagoland water that hydrates fellow
     hyperrhythmic abstract electronic pop groups like
     Mahjongg and Lazer Crystal. This water turns the
     vibraphone of late-90s Chicago post-rock into a synth,
     and adds a phat trigger to every drum. Chandeliers'
     debut album <em>The Thrush</em> was recently released on
      Pickled Egg (UK), and on the group's own Obey Your
     Brain imprint (co-run by Griffin of Icy Demons, also
     home to Man Man and Need New Body). - Jason Sigal 2008</
     span></p>\n<p><span style="font-family:arial, sans-serif
     ;font-size:13px;border-collapse:collapse;"><em>"A  cut
     from the lost Zapp album that was produced by Chris
     Carter, 'Whip the Gift' is ethereal, mournful robo-soul
     at its ghost-in-the-shell finest... it is surely a
     beauteous thing indeed. But it is not all machine-
     grievance, the Chandeliers part of the EP also contains
     the guaranteed boogie-splice euphoria of 'Sugarscoop' -
     which is not leaving 20 JFG's bone encrusted DJ box for
     the foreseeable future." <strong>20jazzfunkgreats music
     blog</strong>. </em></span></p>'],
 'image': ['https://freemusicarchive.org/file/images/artists
     /Chandeliers_-_2009113014608987.png'],
 'tags': ['chandeliers'],
 'unique_id': ['100'],
 'website': ['thechandeliers.com']
 }
```


Figure 4.1: Examples from the web-scraped artist biography corpus, illustrating the heterogeneity of content, structure, and metadata across sources. The first example shows a short, minimally informative biography dominated by self-identification and links. In contrast, the second example contains a long, multi-paragraph description with embedded HTML formatting, stylistic markup, and quoted external reviews. These examples highlight key properties of the corpus: highly variable length and depth, mixed narrative and promotional content, and inconsistent formatting. While this diversity reflects authentic community-authored descriptions, it also introduces noise and sparsity of explicit biographical facts when used for representation learning.

Length and sparsity. The most consequential property of the corpus is that a majority of its artist entries carry no biography at all. Of the 18,912 deduplicated entries, only 8,484 (44.9%) contain biography text. The remaining 10,428 (55.1%) are artists present in the platform exports, and therefore matchable by name, for whom no usable description survived the filtering step described above. We return to this distinction in Chapter 5, since the coverage figures reported there record whether an artist is *present* in the corpus rather than whether the corresponding entry carries text.

Figure 4.2 shows the length distribution across the 8,484 entries that do carry text. On a logarithmic axis the distribution is approximately log-normal with a pronounced right tail. Its mean is 303.3 tokens with a standard deviation of 566.8, but the median of 166 tokens describes the typical entry better, with quartiles at 77 and 349 tokens. The tail extends far beyond the bulk: the 90th percentile is 649 tokens, the 99th is 1,983, and the longest entry runs to 21,082. Text is correspondingly concentrated, with the longest 10% of biographies accounting for 44.0% of all text in the corpus. At the other end, 6.1% of text-bearing entries fall under 20 tokens and 16.2% under 50, lengths at which a biography states little more than a name and a city. Taken together with the empty entries, only about 37.6% of artists in the corpus have a biography of at least 50 tokens.

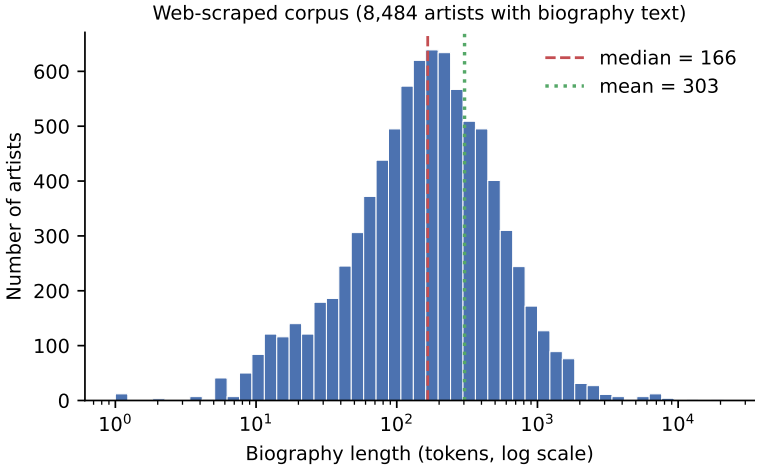


Figure 4.2: Length distribution of the web-scraped biography corpus, restricted to the 8,484 artist entries that carry text; the remaining 10,428 empty entries are excluded from the plot. The horizontal axis is logarithmic. The distribution is heavy-tailed and right-skewed, with the mean (303 tokens) well above the median (166 tokens). Compare with Figure 4.4, whose horizontal axis is linear because the LLM-generated corpus is concentrated almost entirely at a single length.

Variation across sources. Table 4.1 summarises how the three platforms differ. Source attribution is recovered from the platform-specific identifiers and asset URLs retained during extraction; 82 entries (0.4%) could not be attributed and are reported separately.

The Free Music Archive supplies most of the entries (16,881 after deduplication) but is also the sparsest, with fewer than two in five of its artists (39.6%) having a biography, reflecting an open upload platform where writing an artist page is optional. MagnaTune contributes only 730 artists, but every one has a biography and these are the longest on average (443.6 tokens, median 348), as expected of a curated label roster where an editorial description forms part of the catalogue entry. Jamendo sits between the two, describing 84.9% of its 1,219 artists with the widest dispersion of the three sources. Corpus composition and text availability are therefore coupled to the source: the platform contributing 89% of the artist names supplies a much smaller share of the actual text. Because the evaluation datasets draw on these same platforms in different proportions, this coupling propagates into the per-dataset coverage differences reported in Chapter 5.

Source	Artists	With text	Mean	Median	P90	Max
Free Music Archive	16,881	6,686 (39.6%)	265.6	148	568	12,794
MagnaTune	730	730 (100.0%)	443.6	348	844	2,876
Jamendo	1,219	1,035 (84.9%)	448.4	201	1,015	21,082
Unattributed	82	33 (40.2%)	273.2	93	722	2,824
All	**18,912**	**8,484 (44.9%)**	**303.3**	**166**	**649**	**21,082**

Table 4.1: Composition of the web-scraped corpus by source platform. "With text" counts artist entries carrying a non-empty biography after filtering; length statistics, in tokens, are computed over those entries only. The platforms differ substantially in how often a biography exists and how long it is when it does.

Formatting, multiplicity, and language. Formatting adds a further source of noise: 7,703 entries (90.8% of those carrying text) contain embedded HTML markup inherited from platform pages that store descriptions as rendered web content, and the markup is stripped before the text is used as a representation. A further 531 artists are associated with more than one biography, either because they appear on multiple platforms or because a platform stores several descriptions; these are concatenated, with longer passages given priority.

Since the extraction applies no language filter, the corpus is also multilingual. Automatic language identification over the text-bearing entries assigns 90.4% to English and 7.0% to another language across 26 languages, the largest groups being French (3.1%), Spanish (0.9%), German (0.7%), Russian (0.6%), and Italian (0.5%); the remaining 2.6% are too short or too noisy to identify reliably. We retain the non-English entries rather than filtering them, since discarding them would reduce coverage precisely for the independent and non-anglophone artists the corpus is meant to reach. The trade-off is that a text

encoder pretrained overwhelmingly on English will extract a weaker signal from them than their length suggests. At 7% of text-bearing entries this is too small a fraction to alter the conclusions of later chapters, but it contributes to the noise in this representation and compounds the popularity and regional biases discussed in Chapter 7.

4.2.2 LLM-based biographies

While the web-scraped biographies provide high-coverage, community-authored descriptions, they exhibit substantial variability in length, structure, and semantic content. This heterogeneity reflects the open-ended nature of self-authored promotional text, but introduces uncontrolled noise when used as a component of a music representation.

To complement this, we construct a second corpus of artist biographies generated using a large language model (LLM), with the explicit goal of producing *controlled textual representations*. In contrast to web-scraped content, the objective is not to maximise diversity but to standardise style, length, and semantic structure, yielding concise, homogeneous descriptions that encode high-level musical and contextual information.

Corpus construction

We construct this corpus by collecting the set of unique artist names appearing across all datasets in the evaluation pipeline (Chapter 3). Each artist's name is used as the sole input to the generation process, ensuring that the resulting biographies are independent of dataset-specific metadata and reflect only information associated with the artist's identity.

We employ a simple prompting strategy to generate concise, encyclopedic-style biographies (Figure 4.3). We intentionally avoid complex prompt engineering to maintain a consistent and reproducible generation process. All biographies are generated using **gpt3.5-turbo** [37], with a maximum generation length of 128 tokens, a sampling temperature of 0.7, and all other decoding parameters left at their API defaults.

This prompt is intentionally minimal and task-agnostic. It encourages the model to produce general-purpose biographical descriptions without biasing the output toward specific MIR tasks or attributes. Moreover, keeping the prompt simple reduces variability introduced by prompt engineering and improves reproducibility, ensuring that differences observed in downstream experiments can be attributed to the text itself rather than to prompt design choices.

The generation process produces 8,614 biographies, with an average length of 123.9 tokens and a standard deviation of 17.3 tokens. Compared to the web-scraped corpus, whose text-bearing entries exhibit a standard deviation of 566.8 tokens (Section 4.2.1), this indicates a far more consistent length distribution. All generated biographies are in plain text, without formatting or stylistic annotations, and every artist name submitted to the generator returns an entry, so the corpus has no empty records.

```
{
  "system":
    "I am a music expert specialising in music biographies.",
  "user":
    "Generate a 100-word biography of the music artist
     ##ARTIST NAME##:"
}
```

Figure 4.3: Prompt used to generate the LLM biographies

Length distribution. Figure 4.4 shows what this consistency actually consists of, and the shape is not that of a naturally uniform corpus. The distribution is not centred on a typical length but concentrated in a single spike at the generation limit: 7,942 of the 8,614 biographies (92.2%) reach exactly the 128-token cap, and 86.4% end without terminal punctuation, that is, mid-sentence. The homogeneity of this corpus is imposed by the decoding constraint rather than discovered by the model, and the low standard deviation reflects a ceiling rather than agreement about how much there is to say. The truncation visible in the first example of Figure 4.5 is thus the modal case rather than an occasional blemish.

The 672 biographies below the cap are those for which the model genuinely had less to say. Around a third (213) are the refusal responses discussed below, which are short by nature at a median of 51 tokens; the other 459 are brief but genuine descriptions averaging 85.6 tokens, typically for artists the model could identify only partially. The shortest entry runs to 20 tokens.

The length homogeneity of this corpus is therefore real but shallow: it delivers a predictable input size for the encoder, which is what the design was aiming at, but it does not indicate that the biographies contain comparable amounts of information. Because generation also stops at a fixed budget rather than at a natural endpoint, what survives is whatever the model stated first. Since the prompt requests an encyclopedic summary, this favours the opening conventions of that genre, such as origin, formation date, and genre affiliation, over later material such as influences or discography. That is convenient for the Wikidata-grounded evaluation in Section 4.3, whose target properties are drawn from precisely those conventions, and a further reason to read its results as favourable to this corpus by construction.

Initial analysis

As in the web-scraped corpus, we use the artist's stage name as the primary identifier to link biographies to audio samples. The same set of artist names is used across both corpora to maximise overlap. However, in the LLM setting, this design introduces an additional source of ambiguity, as the model must implicitly resolve the artist's identity from the provided name alone.

Figure 4.5 illustrates typical examples from the generated corpus. These

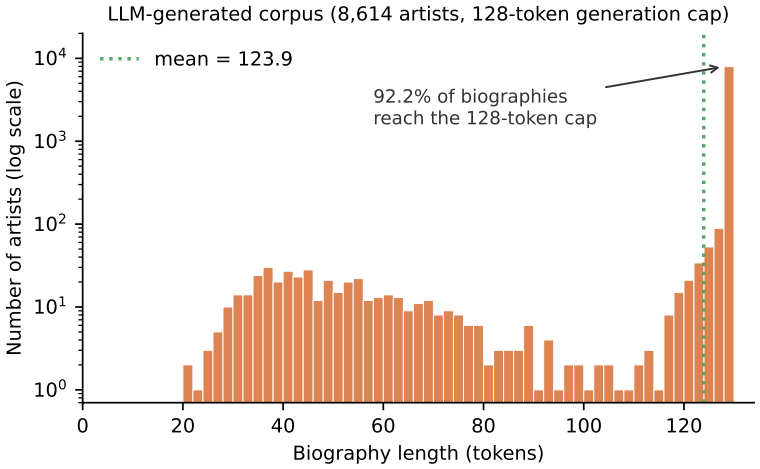


Figure 4.4: Length distribution of the LLM-generated biography corpus. The vertical axis is logarithmic, without which the sub-cap tail would be invisible. Unlike the broad, heavy-tailed web-scraped distribution in Figure 4.2, this corpus is concentrated almost entirely at a single value: 92.2% of biographies reach the 128-token generation cap and are truncated mid-generation. The dispersed lower mass consists of refusal responses and short descriptions for artists the model could only partially identify.

examples highlight three key properties of the LLM-generated biographies.

First, the generated text is highly homogeneous in structure and style. Most biographies follow a concise, encyclopedic format, consistently encoding attributes such as origin, genre, and career milestones. This reflects the intended design of producing controlled textual representations.

Second, the generation process exhibits truncation effects. As quantified above, the 128-token limit is reached by 92.2% of biographies, so ending abruptly is the norm rather than the exception, and the model would clearly have produced longer outputs if unconstrained. This enforces a consistent upper bound on text length, the property the design was aiming for, at the cost of leaving most descriptions grammatically incomplete.

Third, the model occasionally produces refusal responses when the artist name cannot be confidently resolved (e.g., “SMAO”). These cases highlight a useful property of LLM-based generation: instead of hallucinating arbitrary content, the model may abstain when insufficient information is available. We identify refusals by matching biographies containing phrases such as “not a known”, “no information”, or “cannot generate”, yielding 216 refusal cases out of 8,614 generated biographies (2.5%).

Despite these desirable properties, the approach introduces a significant risk of entity ambiguity and hallucination. When multiple artists share the same name, the model tends to default to the most prominent entity, potentially

```
{
  'author_name': 'The Trammps',
  'biography': 'The Trammps were an American disco and soul
      group from Philadelphia, Pennsylvania. Formed in the
      early 1970s, the group consisted of lead singer Jimmy
      Ellis, Harold Doc Wade, Stanley Wade, Robert Upchurch,
      and Earl Young. They were known for their energetic and
       danceable music, featuring strong vocals and funky
      beats. The Trammps achieved great success in the disco
      era with their hit songs "Disco Inferno" and "That\'s
      Where the Happy People Go." They also received a Grammy
       Award for "Best R&B Vocal Performance by a Group" for
      their hit single "Disco Inferno." The Tr'
}

{
  'author_name': 'Eminem',
  'biography': 'Eminem, born Marshall Bruce Mathers III in 1
      972, is an American rapper, songwriter, and record
      producer. He rose to fame in the late 1990s with his
      debut album, "The Slim Shady LP," which won him a
      Grammy Award for Best Rap Album. Eminem's music often
      reflects on his troubled upbringing, struggles with
      addiction, and relationships with women. He is also
      known for his controversial lyrics, which have been
      criticised for their violent and misogynistic content.
      Despite this, Eminem has become one of the most
      successful and influential hip hop artists of all time,
       with over 150 million records'
}

{
  'author_name': 'SMAO',
  'biography': 'Unfortunately, as an AI language model, I
      cannot generate a biography for SMAO as I am not aware
      of any music artist or band by that name. Could you
      provide me with more information or context about SMAO?
       Such as their genre, nationality, or any notable
      achievements or events in their career? I'll be happy
      to assist you in writing a compelling biography.'
}
```

Figure 4.5: Examples from the LLM-generated artist biography corpus, illustrating typical generation behaviours and failure modes. The first two examples show concise, encyclopedic-style summaries that follow a consistent narrative structure and explicitly encode biographical attributes such as origin, genre, and career milestones. The third example illustrates a refusal case, where the model declines to generate a biography due to insufficient information associated with the provided artist name. Across examples, the biographies are homogeneous in style and length and free of formatting noise, but they remain sensitive to name ambiguity and truncation effects due to token limits.

generating incorrect but plausible biographies. This issue is inherent to name-based conditioning and motivates the factual evaluation presented in Section 4.3.

A related concern is the relationship between this corpus and encyclopedic sources such as Wikipedia. Since Wikipedia forms part of the training data of most large language models, the generated biographies of well-documented artists are likely to paraphrase Wikipedia-like content. For prominent artists, then, this corpus is better regarded as a reformatted, uniformly styled projection of encyclopedic knowledge than as an independent source; its distinctive value lies in extending that projection to the long-tail artists for whom no Wikipedia page exists. This overlap also has consequences for any evaluation grounded in Wikipedia-derived resources, which we take up in Section 4.3.

4.3 Factual accuracy analysis

To assess the reliability of the two corpora we have described, we construct a question-answering benchmark to evaluate the factual content encoded in the biographies. To achieve our goal, we leverage WikiData as a source of facets of our artists.

We start by selecting all 4,089 artists that appear in both datasets. From this set, we filter out all artists with empty or non-English biographies, yielding a list of 1,822 artists. We then use the Wikidata API to retrieve the associated Wikidata entity. Based on these results, we filter out any results unrelated to music groups or musical professions. We obtain a corpus with 384 artists' entities. This collection has been publicly released.[6]

Using the retrieved entities, we construct the question-answering corpus using the following attributes of the Wikidata entities, *if present in the objects*:

- **Inception**: What year was this group created?
- **Start of work period**: What year did this artist start to produce music?
- **Country of origin**: From which country is this group from?
- **Place of birth**: Where was this artist born?
- **Genre**: What genre does this artist produce?
- **Instrument**: What instrument does this artist use?
- **Members**: Who are the members of this band?

This process produces a list of 1,164 questions to assess the content of the proposed corpora. We then choose **BLEU** as our evaluation metric.

4.3.1 Question answering evaluation

To quantify how much factual information is encoded in each corpus, we compare the answers extracted from biographies with ground-truth values obtained from Wikidata. This evaluation is not intended to assess linguistic quality, but rather to measure the extent to which biographies contain *explicit, retrievable factual content*.

[6] `https://github.com/fedingo/music_multimodal_augmentations/blob/main/wikidata_artists.json`

For this purpose, we use BLEU as a token-level overlap metric between the extracted answer $\hat{y}$ and the ground-truth answer y. Although BLEU was originally introduced for machine translation, it provides a simple and interpretable proxy for measuring lexical agreement in short text spans:

$$BLEU_w(\hat{y}; y) := \exp\left(\sum_{n=1}^{N} w_n \ln p_n(\hat{y}; y)\right). \tag{4.4}$$

In our setting, answers are typically very short (often single tokens or short phrases), making higher-order n-gram matching unreliable. Therefore, we use a unigram configuration ($w_1 = 1$), effectively reducing BLEU to a precision-based measure of token overlap. This choice reflects the nature of the task: we are primarily interested in whether the correct factual tokens (e.g., country names, years, genres) are present in the extracted answers. Unlike exact match, BLEU-1 also assigns partial credit to approximately correct answers. For example, if the ground-truth place of birth is a specific town in Scotland, but the extracted answer is simply "Scotland", the Exact Match would score this as zero, whereas BLEU-1 captures the partial overlap.

Alternatives considered. Two alternatives were considered before settling on BLEU-1, and the resulting choice has limits that we state alongside them.

The first was **Exact Match**, the conventional metric for short-answer question answering. It was rejected for the reason illustrated above: the quantity we are trying to measure is whether a biography *contains* a given fact, and the granularity at which a biography states a fact rarely matches the granularity at which Wikidata records it. Wikidata stores place of birth as a municipality, whereas a biography typically states a country or region; Wikidata stores an inception year, whereas a biography may say "the early 1970s". Under Exact Match, a biography that correctly locates an artist in Scotland scores identically to one that says nothing at all, which would systematically understate the factual content of exactly the descriptive, narrative register that the web-scraped corpus is written in. Since the comparison of interest is *between* corpora of different registers, a metric that penalises one register categorically would confound the result. BLEU-1 assigns partial credit in these cases and therefore separates "imprecise" from "absent".

The second was **human evaluation**, which would have been the more informative option: a human judge can assess whether an answer is factually correct rather than lexically overlapping, and could distinguish a genuine statement of a fact from a plausible inference by the QA model, a distinction we can only flag qualitatively below. It was rejected on cost. Judging 1,164 questions against three corpora requires roughly 3,500 judgements per annotator, and a defensible protocol would require multiple annotators and an agreement analysis. This was not feasible within the scope of this evaluation, which is one component of a larger experimental programme rather than its main object.

We note that this was a pragmatic choice rather than the outcome of a systematic metric study, and we did not empirically compare BLEU-1 against

other automatic alternatives such as token-level F1, ROUGE, or embedding-based similarity measures. BLEU-1 has known weaknesses in this setting: as a precision-based measure it does not penalise verbose answers, and as a lexical measure it cannot recognise a correct answer expressed in different words (e.g. "UK" for "United Kingdom"). Its role here is therefore deliberately limited. The absolute BLEU-1 values are not meaningful in isolation and we do not interpret them as accuracy; what the evaluation supports is the *relative* comparison across corpora, since all three are scored by the same extractor against the same questions with the same metric. Any residual bias from the metric applies uniformly across the compared corpora, and the qualifications we raise when interpreting the results in Section 4.3.2 are concerned with effects (entity ambiguity, inference by the extractor, evaluation-set bias) that a different automatic metric would not have removed.

Choice of answer extractor. To extract answers from the biographies, we use the UnifiedQA-3B model [24] in an extractive setting. We required a strong, well-established extractive QA model: one that reads a supplied passage and returns a span or short phrase grounded in it, rather than composing an answer from its own knowledge. UnifiedQA is trained across a range of QA formats and was a widely used and reliable choice at the time, and its extractive framing means that when the biography does not contain the requested fact, the model has no supported answer to return.

We also considered using a general-purpose instruction-following LLM (specifically ChatGPT) as the answer extractor, and rejected it. The risk is structural rather than incidental: such a model has extensive parametric knowledge about the very artists we are asking about, and that knowledge is precisely the quantity the protocol is designed to measure in the *text*. Asked "From which country is this group from?" with a biography that never mentions a country, an instruction-following model can answer correctly from memory, and the resulting score would report what the model knows rather than what the biography encodes, collapsing the distinction the entire evaluation rests on. This concern is sharpened by the fact that one of the corpora under evaluation was itself generated by a model of that family. UnifiedQA-3B is smaller and extractively framed, and is therefore substantially more likely to return only what the provided biography supports.

This mitigates, but does not eliminate, the risk of knowledge leakage: as we show in Section 4.3.2, UnifiedQA still occasionally produces a plausible inference ("Scotland" from "Scottish-American", or a default of "guitar" for an unstated instrument) rather than an extracted span. We treat this as a known upper-bound bias affecting all three corpora rather than a resolved problem, and return to it when interpreting the results.

In addition to the two corpora proposed in this chapter, we include MSD-Artist as a reference baseline in the factual evaluation. MSD-Artist is an artist-level metadata resource distributed with the Million Song Dataset, containing short descriptive text associated with artists. Although it is not constructed as

a biography corpus, it provides a useful point of comparison with an existing artist-level textual source.

4.3.2 Experimental results

		Web-scraped	**LLM-based**	**MSD-Artist**
Property	**Count**	BLEU1	BLEU1	BLEU1
Inception	158	24.2	**56.3**	26.4
Start of work	260	17.2	**43.5**	19.2
Country of origin	227	46.8	**61.7**	42.3
Place of birth	79	17.1	**28.5**	15.8
Genre	278	23.4	**36.9**	16.1
Instrument	79	33.7	**38.1**	19.3
Members	83	23.4	**33.7**	13.7
Overall	1164	27.0	**45.1**	23.3

Table 4.2: Evaluation of the factual correctness of the LLM-generated biographies, Web-scraped biographies and MSD-Artist. The MSD-Artist dataset is a collection provided with the Million Song Dataset and is used as a baseline.

Table 4.2 reports the results of the factual evaluation across the web-scraped, LLM-generated, and MSD-Artist corpora. Overall, the LLM-based corpus achieves consistently higher BLEU-1 scores across all properties, with an average improvement of 18.1 BLEU points over the web-scraped corpus. This indicates that LLM-generated biographies contain a higher proportion of Wikidata-aligned factual tokens. However, these results must be interpreted with care, as several factors influence the observed performance.

First, the corpora differ fundamentally in their construction objectives. Web-scraped biographies are user-generated and primarily serve self-promotional purposes. As a result, they often omit canonical biographical facts (e.g., dates, locations, formal career milestones), instead emphasising stylistic descriptions and narrative context. This is reflected in low BLEU scores for properties such as Inception, Start of work period, and Place of birth, where the relevant information is typically absent rather than incorrect. In this setting, low BLEU scores primarily indicate limited factual coverage.

The MSD-Artist corpus provides an additional point of comparison. As a curated metadata resource, it contains concise descriptions focused on high-level stylistic attributes. While it performs comparably on coarse categories such as genre, it lacks detailed factual information, resulting in lower scores for temporal and locational properties. This highlights that curated metadata, while informative, is not designed to provide comprehensive factual grounding.

In contrast, the LLM-generated biographies are explicitly prompted to produce concise, encyclopedic-style summaries. This encourages the inclusion of standard biographical attributes (e.g., origin, genre, membership), which align closely with Wikidata facets. Consequently, the LLM corpus achieves substan-

Question 1: Where was David Byrne born?
Target 1: Dumbarton
Question 2: In what year did David Byrne start to produce music?
Target 2: 1973
Question 3: What instrument does David Byrne use?
Target 3: [Guitar, Voice]

Web Biography:
David Byrne (born May 14, 1952) is a Scottish-American musician and artist perhaps best known as a founding member and principal songwriter of the new wave band Talking Heads, which was active between 1974 and 1991...
Answer 1: Scotland
Answer 2: 1972
Answer 3: Guitar

ChatGPT Biography:
David Byrne is a Scottish-American singer, songwriter, and musician, best known as the frontman of the iconic new wave band Talking Heads. Born in 1952 in Dumbarton, Scotland... In 1975, he formed Talking Heads... Byrne's distinctive vocals, quirky lyrics, and...
Answer 1: Dumbarton, Scotland
Answer 2: 1975
Answer 3: Vocals

Figure 4.6: Question-answering examples for the artist *David Byrne*, illustrating different sources of error and inference in the factual analysis. The web-scraped biography does not explicitly state the artist's place of birth, yet the QA model infers a country-level answer ("Scotland") from the descriptor "Scottish-American". For the Start of the work period, all sources provide slightly different years, highlighting inconsistencies across knowledge sources, even for well-documented artists. Finally, when the instrument is not explicitly mentioned, the QA model defaults to a stereotypical inference ("guitar"). In contrast, the LLM-generated biography explicitly mentions "vocals", leading to a different but still plausible answer.

Question: Where was John Williams born?
Target: The Bronx

Web Bios:
I'm a guitar player, mostly electric, steel string acoustic too. I do session work, songwriting, and production projects. I have a well-outfitted home studio for songwriting and some production, and I also work with a nationally acclaimed pro studio here in Seattle...
UnifiedQA Response: Texas

ChatGPT Bio:
John Williams is a legendary American composer, conductor, and pianist who was born in New York City in 1932. He is widely regarded as one of the greatest film composers of all time, having written scores for over 100 films, including iconic works such as Jaws, Star Wars...
UnifiedQA Response: New York City

MSD-A:
There are at least five artists with this name: (1) John Towner Williams (film score composer), (2) John Christopher Williams (classical guitarist), (3) John Williams (Irish musician), (4)...
UnifiedQA Response: England

Figure 4.7: Question-answering examples for the artist name *John Williams*, illustrating the impact of entity ambiguity on factual evaluation. The web-scraped biography corresponds to a local musician unrelated to the film composer, leading the QA model to produce an incorrect answer. The LLM-generated biography collapses the ambiguous name to the globally prominent film composer, yielding a plausible but mismatched entity. The MSD-Artist metadata explicitly lists multiple artists sharing the same name, producing yet another conflicting answer. The MSD-Artist result is reported as a reference with an existing dataset from Oramas et al. [34].

tially higher scores for these properties, particularly for Inception (+32.1), Start of work period (+26.3), and Country of origin (+14.9). These gains suggest that, for well-identified artists, the model reproduces widely documented facts with reasonable accuracy.

Second, the evaluation set is biased toward well-documented artists. Of the 384 evaluated entities, 312 have corresponding Wikipedia pages. Since Wikipedia content overlaps with LLM training data, this creates favourable conditions for the LLM-based corpus, the leakage effect anticipated in Section 4.2.2. As a result, the observed improvements are partly attributable to alignment between the evaluation source (Wikidata/Wikipedia) and the model's training distribution, limiting generalisability to less prominent artists.

Third, the extractive QA process introduces additional uncertainty. Although UnifiedQA-3B is used to minimise knowledge leakage, it may still infer plausible answers when the biography does not explicitly contain the required information. Figure 4.6 illustrates both cases: the model infers a country of origin from the descriptor "Scottish-American", and defaults to "guitar" where no instrument is stated. In such cases, non-zero BLEU scores may reflect plausible inference rather than explicit factual content.

Finally, name ambiguity remains a critical source of error. As illustrated in Figure 4.7, ambiguous artist names (e.g., "John Williams") can lead to mismatched entities across sources. LLM-generated biographies tend to default to the most prominent entity, while web-scraped biographies may refer to local or unrelated artists. This limitation is inherent to name-based conditioning and highlights the need for caution when using artist names as identifiers.

Overall, this analysis shows that LLM-generated biographies encode denser and more standardised factual content, but this advantage is driven in part by alignment with encyclopedic knowledge sources and by dataset bias. Importantly, higher factual overlap does not necessarily imply greater usefulness for downstream MIR tasks. These findings motivate the evaluation in Chapter 5, where representations are assessed directly on music understanding tasks rather than proxy factual metrics.

RQ1 (Resource) hypothesised that artist-level information is abundant enough online to be aggregated through web scraping and synthesised with large language models, producing corpora that cover the majority of artists in standard MIR datasets (Chapter 1). These results confirm that artist biographies can be scalably constructed using web scraping and large language model generation, achieving consistently high coverage across all evaluation datasets. However, the analysis also reveals that scalability comes with trade-offs the hypothesis did not anticipate: LLM-generated biographies are more homogeneous and factually dense, but susceptible to entity ambiguity, while web-scraped biographies offer greater diversity at the cost of sparsity and noise. The *reliability* component of RQ1 is therefore only partially satisfied, a nuance we return to in Chapter 7.

4.4 Contributions

This chapter makes the following contributions to music representation augmentation in music information retrieval.

Two reproducible artist biography corpora for MIR. We construct and make available two large-scale, artist-level textual corpora:

- A high-coverage collection of web-scraped, community-authored biographies, and
- A corpus of concise, homogeneous biographies generated using a large language model.

Both corpora are aligned with common MIR benchmarks and designed to enable reproducible multimodal research. As described in Section 4.2.1, the web-scraped corpus is distributed as metadata together with the procedure needed to reconstruct its text, since the source platforms' licences do not uniformly permit redistribution of the biographies themselves.

A factual evaluation framework for textual knowledge sources. We propose a Wikidata-grounded evaluation protocol based on extractive question answering to assess the factual content of artist-level text independently of downstream task performance. This framework enables controlled, quantitative comparison of heterogeneous knowledge sources along dimensions such as coverage and factual alignment.

An empirical analysis of community-authored vs LLM-generated knowledge. We show that LLM-generated biographies encode denser, more standardised factual information for well-documented artists, while web-scraped biographies provide broader stylistic diversity but sparser factual content. This analysis highlights the trade-offs between realism, coverage, and factual consistency, and motivates their complementary use in downstream MIR tasks.

4.5 Summary

This chapter introduces and analyses two textual corpora for artist-level knowledge augmentation in music representation learning: a web-scraped corpus of community-authored biographies and a corpus of biographies generated using a large language model. Both are designed to address limitations of existing textual resources in MIR, particularly with respect to coverage, availability, and reproducibility.

Our analysis highlights a fundamental trade-off between these sources. The web-scraped corpus provides broad coverage and reflects the diversity of real-world artist descriptions, but often lacks explicit factual information. In contrast, the LLM-generated corpus produces more homogeneous, encyclopedic-style text that encodes a higher proportion of verifiable facts for well-documented artists, while remaining susceptible to entity ambiguity and hallucinations.

To characterise these differences, we introduce a Wikidata-grounded factual evaluation framework based on extractive question answering. This analysis shows that higher factual alignment does not necessarily imply greater suitability for downstream MIR tasks; rather, it reflects differences in the type of information encoded by each corpus.

Overall, this chapter establishes both the potential and the limitations of using artist-level textual knowledge for music representation learning. These findings motivate the next chapter, where we evaluate how different textual sources interact with audio representations in multimodal settings.

Chapter 5

Comparing music representations

Chapter 3 introduced the evaluation framework for music representations, and Chapter 4 described the construction of two corpora of artist biographies. Building on these foundations, this chapter presents a systematic comparison of audio representations, augmented textual representations, and their multimodal integration. We show that combining audio with aggregated textual representations yields consistent gains across five of the six evaluated tasks and datasets.

The central objective of this chapter is to assess whether incorporating external knowledge provides measurable and consistent benefits across music information retrieval (MIR) tasks. To this end, we evaluate multiple representation variants within a unified experimental framework, enabling controlled comparison across the genre classification, emotion recognition, and music tagging benchmarks.

Our analysis shows that augmented textual representations encode semantic information that is complementary to audio features. Although textual representations alone are often insufficient to match audio-only performance, aggregating and integrating them within a multimodal model improves performance. Through controlled ablation studies, we further examine the extent to which these improvements can be attributed to architectural capacity, metadata usage, or explicit genre cues.

The remainder of this chapter is structured as follows. We first define the evaluated representations and discuss their coverage across datasets. We then present results for individual and aggregated representations, followed by multimodal experiments and comparison with representative models from the literature. Finally, we conduct a series of ablation studies to analyse the contribution of each component of the proposed approach.

5.1 Evaluated representations

In Chapters 3 and 4, we define music representations and their augmentation with external knowledge. We now describe in detail the specific representations evaluated in this chapter.

5.1.1 Audio and metadata music representations

These representations rely solely on the information contained in the multimodal triplet (w, x_t, x_a), and can therefore be expressed as instances of the transformation g.

Audio - Mel spectrogram

As described in Section 3.1, Mel Spectrograms are a well-established method to represent audio in MIR and speech recognition. In this case, the transformation is defined as:

$$g(w, x_t, x_a) = mel_spectrogram(w). \tag{5.1}$$

Text - title and artist

A simple textual baseline is obtained by formatting the title and artist metadata into a fixed template. This is another baseline that allows us to understand the impact of biographies better. The transformation is defined as:

$$g(w, x_t, x_a) = template(x_t, x_a) = \text{“Title: } x_t\text{, Artist: } x_a\text{”}. \tag{5.2}$$

Text - lyrics

This representation has been studied in prior work [13]. While lyrics are typically obtained from external corpora, we instead use a speech recognition system to transcribe them directly from audio. This choice leads to noisier transcription, but allows us to keep this approach reproducible. The system used for these transcriptions is “whisper-large”. The transformation is defined as:

$$g(w, x_t, x_a) = transcribe(w). \tag{5.3}$$

This representation has some sparsity issues, as not all tracks contain vocal content (e.g., instrumental music). We report the coverage for each dataset in our evaluation pipeline in Section 5.1.3.

5.1.2 Artist-biography augmented music representations

These representations leverage an artist-mapping method to expand the representation with a biography of the sample's artist. Each corpus provides artist-level biographical information associated with music samples via the artist-to-biography mapping function $k_b(x_a)$. We define our augmented transformation

as:

$$\hat{g}(w, x_t, x_a, k_b(x_a)) = template\big(k_b(x_a)\big) = \text{“Biography: } k_b(x_a)\text{”}, \tag{5.4}$$

where $k_b(x_a)$ returns the biography associated with artist x_a.

Each of these corpora differs in terms of source, level of curation, and generation process, allowing us to assess the impact of different forms of external knowledge on downstream MIR tasks.

- **KILT.** We use the Wikipedia corpus distributed within the KILT benchmark [38] as a source of biographical information. For each artist, a retrieval system is used to obtain a candidate passage. To mitigate noise introduced by unrelated or ambiguous retrievals, we apply a simple heuristic that checks whether the artist's name appears in the retrieved passage. If the artist is not explicitly mentioned, the passage is discarded, and the representation is treated as missing. On average, biographies retrieved using this approach contain 43 tokens.
- **MSD-Artist.** As a comparison point, we use the MSD-Artist corpus released by Oramas et al. [34], which provides curated Wikipedia biographies aligned with artists in the Million Song Dataset. Artists in our evaluation datasets are mapped to this corpus when possible. The MSD-Artist collection contains biographies for 27,641 artists, and the resulting representations average 26 tokens in length in our experimental setting.
- **Web-scraped Biographies.** We use the community-curated, web-scraped biography corpus introduced in Section 4.2.1. This corpus is derived from publicly available artist descriptions created by online music communities and fans. Unlike Wikipedia-based sources, these biographies often emphasise stylistic traits, influences, and genre descriptors, providing a complementary form of external knowledge.
- **LLM-generated Biographies.** Finally, we evaluate a corpus of artist biographies generated using a large language model, as described in Section 4.2.2. These biographies are generated by prompting the model to produce concise, descriptive summaries of artists, capturing high-level semantic information in a controlled and standardised format.

Together, these four corpora allow us to compare automatically retrieved, community-generated, and model-generated external knowledge sources under a unified representation framework.

5.1.3 Representation coverage

As discussed in Section 4.1, coverage is a key factor when evaluating augmented representations. In this section, we define and report the *representation coverage* of all proposed representations across the evaluated tasks.

We define **representation coverage** as the ratio of samples for which a representation produces a non-empty output:

$$coverage(g) = \frac{1}{N} \sum_{i}^{N} \mathbb{I}\left[g(x_i) \neq \emptyset\right], \tag{5.5}$$

	GTZAN	FMA-S	EmoM	Deezer	MTAT	Jam-MT	Overall
Title-Artist	91.1%	99.8%	99.7%	99.9%	99.9%	99.9%	98.4%
Lyrics	66.0%	63.7%	45.8%	84.0%	43.4%	28.0%	55.2%
KILT	88.2%	20.3%	19.9%	82.2%	28.1%	7.2%	41.0%
MSD-Artist	82.4%	5.0%	7.7%	98.2%	25.0%	0.8%	36.5%
Web-Bio	13.9%	100.0%	93.0%	3.8%	76.8%	84.9%	62.1%
LLM-Bio	90.3%	96.9%	98.4%	99.7%	98.7%	94.5%	96.4%

Table 5.1: Representations coverage for each dataset in the evaluation pipeline.

where $\emptyset$ is defined as the empty representation. In practice, this definition coincides with corpus coverage for augmented representations, while remaining applicable to all representation types considered in this work. We omit audio coverage, as audio is available for all samples by construction (i.e., 100% coverage).

Table 5.1 shows substantial variability in coverage across both datasets and representation types. While title-artist metadata and LLM-generated biographies achieve near-complete coverage by construction, other representations exhibit patterns that depend strongly on dataset composition and the source of external knowledge. These differences are critical for interpretation, as performance must be considered in light of the proportion of samples for which each representation is available.

Coverage for the lyrics representation reflects the presence of vocal content in each dataset. Datasets such as Jamendo-MT and MTAT, which contain a high proportion of instrumental tracks, exhibit particularly low coverage. This highlights that lyrics capture specific properties related to vocal presence rather than providing a general-purpose representation.

Wikipedia-based representations (KILT and MSD-Artist) exhibit similar coverage patterns across datasets, reflecting their shared reliance on encyclopedic sources that emphasise well-known artists. This effect is particularly pronounced for the Deezer dataset, which is derived from the Million Song Dataset and therefore aligns closely with MSD-Artist coverage. More broadly, these representations achieve high coverage on datasets dominated by mainstream artists but perform poorly on collections that emphasise independent or long-tail music.

In contrast, the Web-scraped biography corpus exhibits complementary coverage behaviour relative to Wikipedia-based sources. Datasets such as FMA-Small and EmoMusic, which emphasise independent artists and community-curated collections, show high coverage under the Web-Bio representation while exhibiting limited coverage under Wikipedia-based augmentations. This complementarity highlights how different external knowledge sources reflect distinct musical ecosystems and artist populations.

One qualification applies to the Web-Bio row. Coverage here records whether the sample's artist is *present* in the corpus, and as Section 4.2.1 shows, presence in the web-scraped corpus does not guarantee that the entry carries text. Requiring a non-empty biography lowers the overall figure from 62.1% to 56.5%, and on FMA-Small, where every artist is matched, from 100% to 74.4%; the

other corpora are unaffected, since they return text whenever they return a match.

Finally, LLM-generated biographies provide consistently high coverage across all datasets, independent of artist popularity or corpus overlap. This uniform availability distinguishes them from all other external knowledge sources considered in this thesis and makes them particularly appealing from a scalability perspective. The residual gaps below 100% coverage are not a property of the generation process itself, but an artefact of the batch generation run: they correspond to artists for which the API call repeatedly failed to complete (transient service errors, timeouts, and rate-limit failures encountered while iterating over the artist list). Because the generation procedure is conditioned only on the artist name, and every evaluation sample carries one, these gaps could in principle be closed simply by re-issuing the failed calls. We report the corpus as it was generated and released, rather than retrospectively patching it, so that the coverage figures correspond exactly to the publicly distributed resource.

It is worth being precise about what *is* counted as covered here. The refusal responses characterised in Section 4.2.2 (216 of 8,614 biographies, 2.5%) are retained in the corpus and therefore count towards coverage, since they produce a non-empty representation. This is a deliberate choice: a refusal is an honest signal that the model could not resolve the artist, and discarding or overwriting it would either lower coverage artificially or replace an abstention with hallucinated content. It does, however, mean that the LLM-Bio coverage figures in Table 5.1 should be read as an upper bound on *informative* coverage, with a small fraction of entries carrying no biographical content. Conversely, the same generation procedure offers a natural gap-filling mechanism for the *other* corpora: an LLM-generated biography can stand in wherever a web-scraped or encyclopedic biography is missing. The combined representation introduced below realises a simple version of this idea by aggregating all available sources per sample. We discuss the limitations of LLM-generated biographies in Chapter 7.

Taken together, these results show that coverage limitations are not random but reflect systematic biases in both dataset composition and external knowledge sources. As a result, relying on any single augmented representation introduces dataset-dependent coverage gaps. To mitigate this effect, we introduce a combined representation that aggregates all available textual sources for each sample, increasing effective coverage and reducing sensitivity to the choice of external corpus. This combined representation is described in the following section and evaluated alongside individual representations in subsequent experiments.

5.1.4 Text-combined music representation

Finally, we define a combined textual representation that aggregates all previously introduced textual sources into a single input. Each component is prefixed with a source identifier (e.g., “Title:” or “Lyrics:”), allowing the model to distinguish between different types of information while leveraging them jointly.

To define this representation formally, let $\mathcal{G}$ denote the set of textual representation functions considered in this chapter (title-artist, lyrics, and all biography-based augmentations). We define the combined textual representation as:

$$g_{\mathrm{c}}(w, x_t, x_a) = \mathrm{concat}\left(\{\mathrm{tag}(r, r(w, x_t, x_a)) \mid r \in \mathcal{G}\}\right), \tag{5.6}$$

where $\mathrm{tag}(\cdot)$ prepends a source identifier (e.g., ``Wikipedia:'') to each representation, and $\mathrm{concat}(\cdot)$ concatenates the resulting strings using a fixed deterministic ordering.

A key consequence of this aggregation is that the resulting representation can become substantially longer than individual inputs. As discussed in Section 5.2.1, the text encoder operates with a maximum context length of 512 tokens. When this limit is exceeded, the input is truncated to the maximum allowable length.

In practice, truncation affects approximately 4% of samples overall, increasing to 8.5% for the FMA-Small dataset. This indicates that, while most samples fit within the context window, aggressive aggregation can introduce information loss in a non-negligible subset of cases.

This representation provides a simple yet effective mechanism to maximise textual coverage while preserving source-specific signals, and serves as the primary textual input in the multimodal experiments presented in the following sections.

5.2 Experimental setup

This section describes the experimental setup used to evaluate the representations introduced in Section 5.1. We detail the model architecture, training objectives, and optimisation settings used across all experiments. All datasets, splits, and evaluation metrics follow the protocol defined in Chapter 3.

Figure 5.1 illustrates the multimodal architecture used in all experiments. Only the input representation varies across experiments, while the encoder and fusion components remain fixed.

5.2.1 Model architecture

All experiments are conducted using a unified multimodal architecture composed of modality-specific encoders followed by a lightweight multimodal fusion module. This design allows us to isolate the impact of different input representations while keeping the downstream model architecture fixed across experiments.

The design intent constrains every choice that follows, so we state it first. The goal of this chapter is not to propose a new multimodal architecture, but to build an instrument for testing the representation hypothesis: a model that accepts any of the representations defined in Section 5.1, treats them all identically, and is simple enough that observed differences can be attributed to the input rather than to the model. The architecture is therefore deliberately minimal, and each component is selected for stability and reproducibility rather than for

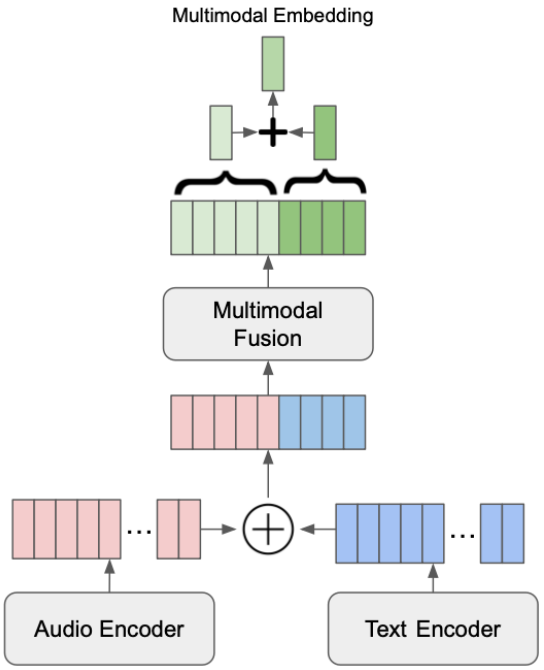


Figure 5.1: Overview of the multimodal architecture used in all experiments. Audio and textual inputs are encoded using modality-specific encoders, producing variable-length sequences of hidden states. These representations are concatenated and processed by a Transformer-based multimodal fusion module. The fused representations are then aggregated into a fixed-dimensional multimodal embedding used by task-specific prediction heads.

peak performance. The one component that required non-trivial design effort is the fusion step, for the reasons described below.

Audio encoder. As the audio encoder, we use VGGish [21], a convolutional architecture operating on Mel Spectrogram inputs and pretrained on AudioSet [19], a large-scale dataset annotated with a rich ontology of sound events.

Given an audio waveform, the encoder produces a variable-length sequence of 128-dimensional embeddings. These embeddings are projected to 1024 dimensions via a linear layer to match the dimensionality of the textual representations. The audio encoder is fine-tuned end-to-end during training as part of the model.

Text encoder. All textual inputs are encoded using the encoder component of T5-Large [42], a Transformer-based model pretrained in a multi-task text-to-text setting.

The encoder produces a variable-length sequence of 1024-dimensional contextualised embeddings. The same text encoder is used across all textual representations to ensure comparability. The text encoder is fine-tuned jointly with the rest of the model.

VGGish and the T5-Large encoder were selected as strong, widely adopted encoders for their respective modalities at the time these experiments were conducted, and both are publicly available with released weights, which keeps the pipeline reproducible. We note that neither encoder is the object of study: as argued in Chapter 3, holding the encoder fixed is what allows the representation to be isolated as the experimental variable. Section 5.5 returns to this point empirically, showing that VGGish is competitive with stronger audio-only models from the literature and that the reported gains are therefore not an artefact of a weak acoustic front-end.

Multimodal encoder. To combine audio and textual representations, we use a randomly initialised Transformer-based fusion module composed of a single self-attention layer. The purpose of this layer is not to add modelling capacity, but to make each modality's embeddings contextually aware of the other before they are pooled: a single layer is sufficient for tokens of one modality to attend to the other, while keeping the number of randomly initialised parameters small relative to the two pretrained encoders.

Modal-specific hidden states are concatenated along the sequence dimension and projected into a shared 1024-dimensional space. Representations are then averaged *within each modality* before the two modality-level vectors are combined into a single multimodal embedding, which is passed to the task-specific prediction heads.

Balancing signal strength across two variable-length modalities. This last detail is the one substantive design decision in the architecture, and it merits explanation because the naive alternative fails in a specific and instructive way. Both modalities produce *variable-length* sequences: the audio encoder emits one frame embedding per short window of the waveform, so its sequence length is a function of track duration, while the text encoder emits one embedding per token, so its length is a function of how much biographical text happens to be available for that artist. Crucially, the two lengths vary independently of each other and independently of the task.

Our initial implementation pooled the concatenated sequence with a single average over all positions, irrespective of modality. Under that scheme, the contribution of a modality to the pooled embedding is proportional to its share of the total sequence length. A track with a long biography and a short audio excerpt is dominated by text; a long instrumental track with a missing biography is dominated by audio. The effective mixing weight of the two modalities is therefore not a property of the model or the task, but an accident of the input lengths, and it changes from sample to sample within a single batch. In practice this made training unstable and caused the model to converge to solutions

that relied almost entirely on whichever modality carried the stronger signal on average for that dataset, rather than learning to use both.

We experimented with several normalisation strategies to correct this imbalance, including rescaling the modality contributions by sequence length and normalising the hidden states before pooling. The approach that proved most stable, and the one used in all reported experiments, is the simplest: compute a fixed-dimensional representation *per modality* first, by averaging within each modality independently, and only then combine the two modality-level vectors. This decouples a modality's influence on the fused embedding from its sequence length, so that a one-sentence biography and a five-paragraph biography enter the fusion step on equal footing with the audio, and each modality is given an equal opportunity to contribute. It also has a practical benefit specific to this study. The textual representations under comparison differ substantially in typical length, from 26 tokens on average for MSD-Artist to several hundred for the web-scraped corpus (Section 5.1), so length-independent pooling removes a confound that would otherwise vary systematically across the very conditions we are comparing.

5.2.2 Training objectives

All models are trained end-to-end using task-specific loss functions. Genre classification is optimised using categorical cross-entropy. Emotion recognition is formulated as a regression task and trained using mean squared error. Music tagging is treated as a multi-label classification problem and optimised using binary cross-entropy loss.

All representations are trained using identical task formulations and loss functions to ensure fair comparison across representation types.

5.2.3 Training and inference settings

All models are trained using the Adam optimiser with default parameters and a batch size of 8, due to GPU memory constraints. Models are trained for 10 epochs on all datasets, except GTZAN and EmoMusic, which are trained for 50 epochs due to their smaller sizes.

Early stopping is applied based on validation performance using the primary evaluation metric for each task. For music tagging, mean Average Precision is used as the early stopping criterion. All reported results are averaged over three independent runs with different random seeds, with the best validation checkpoint selected for test evaluation in each run. Mean and standard deviation are reported for audio, text, and multimodal representations to indicate result reliability.

Handling missing representations and input length constraints. In multimodal experiments, modalities are never explicitly removed. When a textual representation is unavailable for a given sample, the model receives an

empty string as input. This design ensures that all experiments evaluate the intrinsic properties of the representations rather than conditional model behaviour based on input availability.

In unimodal experiments (audio-only or text-only), the same architecture is used with only the corresponding encoder activated. This allows us to keep the architecture fixed across all experiments while varying only the input representation.

Concatenating multiple textual representations can result in long input sequences. As discussed in Section 5.2.1, the maximum context length supported by the text encoder is limited to 512 tokens. When this limit is exceeded, the input representation is truncated to the maximum allowable length.

5.2.4 Evaluation protocol summary

We evaluate all representations on three MIR tasks: genre classification, emotion recognition, and music tagging. Each task is assessed on two benchmark datasets using the splits and metrics defined in Chapter 3. For clarity, we restate the label spaces involved. Genre classification is a single-label problem over 10 balanced genres for GTZAN and 8 balanced genres for FMA-Small. Emotion recognition is a regression problem over the two continuous valence–arousal dimensions for both EmoMusic and Deezer. Music tagging is a multi-label problem over the top 50 tags for MagnaTagATune and 56 mood/theme tags for Jamendo-MoodTheme. Performance is reported using accuracy for genre classification, R^2 for emotion recognition, and ROC-AUC and mean Average Precision for music tagging.

5.3 Performance of representations

In this section, we evaluate each representation in isolation. We consider only single-modality inputs, without multimodal fusion, to assess the relative strengths and limitations of individual representations. This analysis also quantifies the extent to which aggregating complementary textual sources improves performance before introducing multimodal models.

Table 5.2 reports the performance of each representation in isolation, alongside a popularity-based baseline (*Popular*). This baseline is input-agnostic: it is fit on the training split of each dataset and produces the same prediction for every test sample. For genre classification, it predicts the most frequent training label, so its accuracy corresponds to the prevalence of the majority class ($\approx 1/C$ for the balanced C-class datasets used here). For emotion recognition, it predicts the mean valence and arousal values of the training split, which by construction yields $R^2 \approx 0$ up to distribution shift between splits. For music tagging, it assigns each tag a constant score equal to its training-set prevalence; this produces a ROC-AUC of 50.0 (no ranking information) and a mAP equal to the average tag prevalence. The baseline is evaluated on the same test splits and with the same metrics as all other representations, and provides a lower

	GTZAN	**FMA-S**	**EmoM**	**Deezer**
	Acc.	Acc.	R^2	R^2
Popular	10.7	12.5	-1.8	0.0
Audio and Metadata Music Representations				
Audio	**77.2 ± 0.4**	52.7 ± 0.5	**51.4 ± 1.9**	**19.0 ± 0.1**
Title-Artist	40.1 ± 3.5	23.7 ± 2.9	-9.1 ± 3.3	1.0 ± 0.5
Lyrics	29.7 ± 0.7	29.0 ± 1.1	-6.1 ± 1.5	3.4 ± 0.7
Artist-Biography Augmented Music Representations				
KILT	66.8 ± 1.9	19.6 ± 0.8	-5.9 ± 0.7	3.0 ± 0.2
MSD-Artist	44.1 ± 1.5	14.6 ± 1.2	-4.0 ± 4.2	7.3 ± 0.5
Web-Bio	13.1 ± 0.5	40.5 ± 1.9	-2.3 ± 6.4	-0.3 ± 0.4
LLM-Bio	73.1 ± 1.2	53.1 ± 0.9	-0.7 ± 5.1	11.8 ± 0.5
Text-Combined	**77.2 ± 1.2**	**67.6 ± 1.0**	2.8 ± 3.5	15.9 ± 0.6

	MTAT		**Jam-MT**	
	AUC	mAP	AUC	mAP
Popular	50.0	6.5	50.0	3.2
Audio and Metadata Music Representations				
Audio	**89.6 ± 0.0**	**36.8 ± 0.0**	**75.3 ± 0.1**	**13.0 ± 0.0**
Title-Artist	80.1 ± 0.3	23.6 ± 0.2	65.6 ± 0.6	9.0 ± 0.3
Lyrics	72.5 ± 0.0	17.0 ± 0.1	55.1 ± 0.2	4.4 ± 0.4
Artist-Biography Augmented Music Representations				
KILT	62.1 ± 0.1	12.9 ± 0.2	50.5 ± 0.6	3.4 ± 0.0
MSD-Artist	60.6 ± 0.3	12.4 ± 0.1	49.8 ± 0.5	3.5 ± 0.1
Web-Bio	77.5 ± 0.1	21.7 ± 0.2	57.9 ± 0.7	5.7 ± 0.0
LLM-Bio	78.8 ± 0.2	23.1 ± 0.2	56.3 ± 0.5	4.7 ± 0.2
Text-Combined	84.4 ± 0.1	27.5 ± 0.3	68.3 ± 0.7	9.9 ± 0.3

Table 5.2: Evaluation of the performance of each single proposed representation. **Bold** indicates highest score. Multiple representations outperform the *popularity*-based baseline, indicating that the representations contain useful information for the model. Moreover, the combination of all representations *text-combined* shows notable improvements in all evaluated datasets except for EmoMusic, where the score is $R^2 \approx 0$. A higher score is better for all measures.

bound for assessing whether a representation encodes meaningful information. For instance, KILT and MSD-Artist achieve performance close to the baseline on Jamendo-MoodTheme, which is consistent with their low coverage on this dataset (Table 5.1).

Within the "Audio and Metadata" group, both *title-artist* and *lyrics* representations perform reasonably well on genre classification and music tagging, but show limited effectiveness in emotion recognition, where performance remains close to the baseline.

The *lyrics* representation outperforms *title-artist* on FMA-Small and Deezer, but this advantage is largely driven by dataset characteristics rather than the representation itself. FMA-Small contains more clearly separated classes (e.g., "instrumental"), where the presence or absence of lyrics becomes informative. In contrast, GTZAN includes genres such as "classical" and "jazz", where vocal presence is less discriminative. Similarly, Deezer exhibits higher lyrics coverage (Table 5.1) due to construction bias, increasing the utility of transcription-based features.

In music tagging, none of the augmented representations matches the performance of *title-artist*. This can be attributed to two factors. First, music tagging is inherently challenging, as each track may be associated with a diverse and heterogeneous set of labels. Second, all augmented representations operate at the artist level, making different tracks from the same artist indistinguishable. This limits the model's ability to capture song-specific attributes required for fine-grained tagging.

LLM-generated biographies consistently outperform other textual representations on GTZAN, FMA-Small, and Deezer, with particularly strong results in genre classification, where performance approaches that of the audio baseline. This reflects the strong correlation between artist identity and musical style. Further analysis reveals that 45.7% of samples explicitly contain the genre label in the representation, increasing to 59.4% for GTZAN. This suggests that explicit semantic cues drive part of the observed performance. We investigate this effect in more detail in Section 5.5.

Overall, most representations outperform the popularity baseline, indicating that they encode useful semantic information. The main exception is EmoMusic, where all representations yield negative performance. This reflects the difficulty of learning emotional cues in low-resource settings. With only 619 training samples, EmoMusic provides insufficient signal to learn reliable valence–arousal predictions. Consistent with this instability, EmoMusic also exhibits the highest variance across runs, particularly for biography-based representations (e.g., Web-Bio ± 6.4, LLM-Bio ± 5.1), further indicating that results on this dataset should be interpreted with caution.

In contrast, the larger Deezer dataset (11,267 training samples) enables the model to learn from textual representations, resulting in positive R^2 scores. This suggests that textual emotion cues are present but statistically weak, and their usefulness depends strongly on dataset scale. Additionally, Deezer annotations are derived from tag-based metadata, which aligns naturally with textual descriptors. This may further explain the improved performance compared to

EmoMusic, which obtains labels from crowdsourced perceptual ratings.

The negative performance of the *popularity* baseline on EmoMusic is also noteworthy. Under the definition of R^2, predicting the mean yields $R^2 = 0$. However, the baseline computes the mean on the training split and evaluates it on the test split. The resulting negative score, therefore, reflects a distribution shift between splits, which is plausible given the small dataset size.

The *text-combined* representation consistently improves performance across tasks, outperforming all individual textual representations. This demonstrates that aggregating complementary sources increases the amount of usable semantic information and mitigates coverage limitations. We use this representation in the multimodal experiments to assess its contribution when combined with audio features.

These improvements show that textual representations capture meaningful high-level semantic information. Next, we explore how these textual signals can be integrated with acoustic features to evaluate whether their combination yields stronger representations within a unified multimodal framework.

5.4 Multimodal fusion of audio and text representations

The previous sections showed that audio and aggregated textual representations exhibit positive performance across tasks and datasets. We now evaluate a multimodal representation (*MM-Combined*) that jointly leverages audio and the aggregated textual representation (*Text-Combined*) introduced in Section 5.1.4.

	GTZAN	**FMA-S**	**EmoM**	**Deezer**
	Acc.	Acc.	R^2	R^2
Audio	77.2 ± 0.4	52.7 ± 0.5	**51.4 ± 1.9**	19.0 ± 0.1
Text-Combined	77.2 ± 1.2	67.6 ± 1.0	2.8 ± 3.5	15.9 ± 0.6
MM-Combined	**89.3 ± 2.6**	**72.9 ± 0.9**	46.6 ± 2.2	**24.4 ± 0.5**

	MTAT		**Jam-MT**	
	AUC	mAP	AUC	mAP
Audio	89.6 ± 0.0	36.8 ± 0.0	**75.3 ± 0.1**	13.0 ± 0.0
Text-Combined	84.4 ± 0.1	27.5 ± 0.3	68.3 ± 0.7	9.9 ± 0.3
MM-Combined	**90.5 ± 0.1**	**38.3 ± 0.3**	75.1 ± 0.2	**14.8 ± 0.7**

Table 5.3: Experimental results comparing single-modality representations to a multimodal combination of the results. We notice that the multimodal combination shows positive improvements.

Table 5.3 shows that multimodal fusion consistently improves performance over both audio-only and text-only representations across nearly all evaluated datasets. This consistent improvement indicates that audio and textual representations capture complementary aspects of the music signal: audio encodes

low-level acoustic structure. In contrast, textual representations provide high-level semantic and contextual information that is not directly observable from the waveform.

In the case of EmoMusic, textual representations exhibit near-zero predictive power, resulting in a multimodal performance slightly lower than the audio-only baseline (-4.8% relative difference). This behaviour is consistent with the small-data regime discussed previously, where weak correlations between textual descriptions and emotional annotations cannot be reliably learned.

Genre classification benefits substantially from multimodal fusion, as both audio and textual representations are individually strong and encode complementary information. A similar pattern is observed in the Deezer emotion dataset, where textual representations become informative at scale, and multimodal integration yields additional gains over each modality in isolation.

For music tagging, multimodal fusion consistently improves over the audio-only baseline on both MagnaTagATune and Jamendo-MT. Since model selection and early stopping are performed using mean Average Precision (mAP), improvements are more pronounced in mAP than in ROC-AUC.

Overall, these results demonstrate that integrating audio and textual representations yields consistently stronger performance than single-modality approaches across tasks and datasets.

While the previous analysis establishes the effectiveness of our representation framework, we now contextualise these results with respect to existing state-of-the-art systems. We therefore compare our multimodal approach against representative multimodal models from the literature for each task to assess its competitiveness under standard benchmarking settings.

5.4.1 Comparison with models from the literature

In the previous section, we showed how our proposed textual representations can be effectively combined with an audio representation to achieve improved performance on downstream tasks. In this section, we contextualise the performance of our multimodal approach with respect to representative multimodal models from the literature.

Why prior systems are not re-run. The numbers reported for prior work in Tables 5.4 and 5.5 are transcribed from the corresponding publications rather than reproduced by us, and the tables are consequently sparse: each system is reported only on the benchmarks its authors evaluated on. Obtaining each system and running it on the full six-dataset suite under our protocol would fill the empty cells and remove any residual protocol mismatch, but this was not possible here. For most of these systems no public code or trained weights were available at the time of writing, and reimplementing from a published description would introduce exactly the kind of uncontrolled variation this thesis is designed to avoid, since a reimplementation cannot be distinguished from a subtly incorrect one without the original artefact to compare against. Several of the strongest systems additionally rely on pretraining corpora that are

	GTZAN	**FMA-S**	**EmoM**	**Deezer**
	Acc.	Acc.	R^2	R^2
MuLaP* [32]		61.1	**58.5**	
Delbouys [13]				22.5
Discogs [2]		60.9		
MM-Combined	**89.3**	**72.9**	46.6	**24.4**

	MTAT		**Jam-MT**	
	AUC	mAP	AUC	mAP
MuLaP* [32]	89.3 †	40.2 †	76.1	
Delbouys [13]				
Discogs [2]	**90.8**	**38.6**	**76.3**	
MM-Combined	90.5	38.3	75.1	**14.8**

Table 5.4: Comparison with representative multimodal models from the literature. Bold indicates the best-performing method for each dataset and metric. † denotes results that are not directly comparable due to evaluation on a cleaned version of MTAT. * indicates models pretrained on external corpora. Reported results for prior work are taken from the corresponding publications and may follow slightly different experimental settings; they are included to contextualise performance rather than to claim strict protocol equivalence. Higher values indicate better performance for all metrics.

not publicly distributable, marked with * in both tables; in one case we contacted the authors to ask whether their resource could be shared for a controlled re-evaluation, and they were unable to do so because the corpus had been assembled with a commercial partner and had since become proprietary.

We therefore retain the published numbers and treat both tables as contextualisation rather than as controlled comparison. They establish that our results sit in the same performance regime as the published state of the art, which is the claim we draw from them, and small differences between rows should not be read as meaningful. The controlled evidence for the effect of the representation itself, which is the object of this chapter, comes from the ablations in Section 5.5, where every configuration is run by us under an identical protocol.

We report in Table 5.4 a comparison between our multimodal model and representative systems from the literature. Overall, our approach achieves competitive and, in several cases, state-of-the-art performance across the evaluated benchmarks. In particular, our model matches or exceeds the best reported performance among the compared methods on GTZAN, FMA-Small, and Deezer, while obtaining results close to prior work on both music tagging datasets. EmoMusic remains the main outlier, consistent with the analysis in the previous sections, which showed that textual representations provide a limited signal in this small-data regime.

A common trend across the compared systems is the central role of *knowledge injection*. Existing multimodal approaches incorporate external informa-

tion either through large-scale pretraining on music-related corpora [2, 32] or by augmenting inputs with additional contextual signals [13]. Our approach follows a similar direction, but avoids the use of copyrighted *textual* sources, making the knowledge augmentation component more accessible and reproducible.

A limitation of this design is that the effectiveness of augmentation depends on the availability and quality of artist-level information. In particular, coverage and reliability may degrade for long-tail or emerging artists (e.g., sparse community biographies or limited encyclopedic presence), thereby constraining scalability in real-world settings.

Having established its competitiveness relative to prior systems, we next analyse our model's behaviour in more detail. We first contextualise our choice of audio encoder relative to stronger audio-only alternatives to disentangle representation gains from architectural capacity. We then investigate the contribution of artist-biography augmentation through controlled comparisons of different knowledge sources. Finally, we analyse the role of explicit semantic cues in genre classification through targeted masking experiments.

5.5 Ablation studies

The previous sections established that combining audio and textual representations yields consistent performance improvements across multiple tasks and datasets. However, these improvements alone do not explain the mechanisms underlying the gains. In this section, we conduct a series of ablation and control experiments to disentangle the contributions of model capacity, artist-biography augmentation, and explicit semantic cues within the textual representations.

First, we analyse the impact of the audio encoder to determine whether the observed improvements can be attributed to representation design rather than architectural strength. Second, we evaluate the role of artist-biography augmentations by comparing them with a more limited combination of our proposed textual representations within the multimodal framework. Finally, we investigate the extent to which explicit genre-related information drives performance gains by introducing controlled masking experiments in the genre classification task.

5.5.1 Impact of the audio encoder

In this first analysis, we contextualise our choice of VGGish as the audio encoder by comparing it with representative audio-focused models from the literature.

We report in Table 5.5 the results of our evaluation. As in Table 5.4, results for prior work are taken from the corresponding publications rather than reproduced here, for the reasons given in Section 5.4.1, and empty cells are benchmarks the original authors did not report on. The comparison is used to establish that VGGish is not anomalously weak, rather than to rank encoders precisely, and the argument below rests only on the gap being small.

	GTZAN	FMA-S	EmoM	Deezer
	Acc.	Acc.	R^2	R^2
Musicset* **[33]**	**83.5**		64.6	
Calm* **[7]**	79.7		**66.9**	
Lee* **[27]**	72.0			
Delbouys [13]				**20.7**
S3T* **[55]**	81.1	**56.4**		
Weninger [52]			60.2	
VGGish	77.2	52.7	51.4	19.0

	MTAT		Jam-MT	
	AUC	mAP	AUC	mAP
Musicset* **[33]**	**91.7**	41.3	**78.6**	**16.1**
Calm* **[7]**	91.5	**41.4**		
Lee* **[27]**	90.2			
Delbouys [13]				
S3T* **[55]**	89.9	40.9		
Weninger [52]				
VGGish	89.6	36.8	75.3	13.0

Table 5.5: Evaluation of audio-focused models in literature compared with VGGish. This table shows that VGGish is a valid choice as our baseline, demonstrating its competitiveness. * indicates models pretrained on an external corpus. **Bold** indicates best performance reported. A higher score is better for all measures.

Across the evaluated benchmarks, VGGish performs competitively with representative audio-focused models from the literature, typically within a few percentage points of reported results. The largest discrepancy is observed on EmoMusic, where models trained with supervised pretraining achieve substantially stronger results. This behaviour is expected in low-resource settings, where large-scale pretraining provides a significant advantage.

Excluding EmoMusic, the performance gap narrows considerably, supporting VGGish as a strong baseline. These results indicate that the improvements observed in our multimodal experiments cannot be attributed solely to an unusually weak audio encoder. Moreover, VGGish is widely used in prior work [5, 2], further supporting its suitability as a reference model.

5.5.2 Impact of artist-biography augmentations

In this section, we evaluate the contribution of artist-biography augmentations to the overall architecture. To this end, we evaluate them against a more constrained combination of our proposed representations that uses only the *audio and metadata* representations (A&M).

We report the results of these experiments in Table 5.6. These results demon-

	GTZAN	**FMA-S**	**EmoM**	**Deezer**
	Acc.	Acc.	R^2	R^2
MM-Audio&Metadata	73.1	46.1	46.1	19.6
MM-Combined	**89.3**	**72.9**	**46.6**	**24.4**

	MTAT		**Jam-MT**	
	AUC	mAP	AUC	mAP
MM-Audio&Metadata	**90.5**	37.7	74.7	13.4
MM-Combined	**90.5**	**38.3**	**75.1**	**14.8**

Table 5.6: Evaluation of the impact of artists' biographies augmentations compared to using audio and metadata representations, showing consistent improvements across all evaluated datasets. MM-Audio&Metadata includes audio, title-artist and lyrics representations. All experiments are multimodal experiments.

strate that the proposed augmentations yield consistent improvements across all evaluated datasets. In particular, genre classification shows the largest improvement, with an average increase of 21.5%. The substantial gains in genre classification are consistent with the fact that artist biographies frequently encode stylistic and genre-related descriptors, providing high-level semantic cues aligned with the prediction target. We will investigate this in more depth in the next section.

In music tagging and emotion recognition, the impact is more moderate. The main exception is the Deezer dataset, where artist-biography augmentation yields a 4.8% improvement over the baseline. This behaviour is consistent with the construction of the Deezer dataset, in which emotion annotations are derived from tag-based metadata, thereby increasing the alignment between textual descriptors and target labels. We also observe a notable improvement in mAP on the Jamendo-MoodTheme dataset, with an absolute gain of 1.4% (relative 10.4%). This suggests that textual representations encode weak emotional signals that can be exploited when sufficient data is available.

5.5.3 Masking in genre classification

In this section, we investigate the effect of explicit genre mentions on genre classification performance. As we reported in Section 5.3, we find that the target genre is frequently mentioned explicitly in the artist's biographies we use for our augmentations. Specifically, we find that explicit mentions of the target genre appear in 59.4% of samples in GTZAN and 45.7% in FMA-Small.

To assess the impact of these mentions, we mask occurrences of the ground-truth genre label in the biography-augmented representations using a $MASK$ token. We evaluate three setups: the LLM-generated biography in isolation (which contains the most genre mentions), our *text-combined* representation, and the multimodal setup that combines all our proposed representations.

We report in Table 5.7 the results of this experiment. What we find aligns

	GTZAN	FMA-Small
LLM-Bio	73.1	53.1
LLM-Bio + Mask	61.4	52.9
Text-Combined	77.2	67.6
Text-Combined + Mask	72.4	66.0
MM-Combined	89.3	72.9
MM-Combined + Mask	78.3	71.1

Table 5.7: Experimental results when masking mentions of the ground-truth labels from the proposed biography-augmented representations. These results show that masking has a greater impact on smaller-scale datasets (GTZAN - 1000 samples) than on larger-scale datasets (FMA-Small - 8000 samples).

with some of our previous findings. Masking explicit genre mentions produces a substantial performance drop on GTZAN (-11.0 absolute points in the multimodal setting, -9.2 on average across experiments), whereas the impact on FMA-Small is comparatively modest (-1.2 absolute points).

These results indicate that explicit genre mentions contribute to performance, particularly in smaller datasets, but do not fully account for the improvements introduced by biography augmentation. The model continues to benefit from implicit stylistic and contextual information present in the biographies. Importantly, despite the performance decrease, the masked models remain competitive with strong baselines reported in the literature.

Taken together, these ablation studies demonstrate that the improvements observed in Section 5.5.2 cannot be attributed solely to stronger audio encoders, trivial metadata augmentation, or label leakage. Instead, they provide strong evidence that artist-biography augmentation introduces complementary semantic information, thereby enhancing multimodal representations.

These findings are consistent with the hypothesis that artist biographies provide complementary semantic cues that improve performance across MIR tasks. As predicted, the gains are strongest for genre classification, where artist-level identity aligns closely with stylistic categories, and less consistent for affective tasks, where the connection between artist-level descriptions and track-level emotional content is weaker. The ablation studies further clarify that biographies contribute through both explicit lexical cues and broader contextual information, with their relative contributions varying by dataset scale.

5.6 Contributions

This chapter introduces a multimodal framework for integrating music representations with artist-level textual knowledge. The main contributions are as follows:

- We formalise multiple music representations within a unified experimental framework, enabling controlled comparison across three MIR tasks and six

benchmark datasets.

- We show that aggregating artist-level textual augmentations provides complementary semantic information that consistently improves multimodal representations across genre classification, emotion recognition, and music tagging.
- Through ablation studies, we demonstrate that these improvements cannot be attributed solely to stronger audio encoders, generic metadata, or explicit genre mentions within biographies. However, the magnitude of the lexical contribution varies by dataset: on GTZAN, explicit genre terms account for a substantial portion of the gain, whereas on FMA-Small the effect of masking is minimal, suggesting that broader contextual semantics play a larger role in higher-data settings.

The work presented in this chapter has been published in [43].

5.7 Summary

In this chapter, we present a systematic evaluation of music representations augmented with textual artist biographies and their multimodal combination across three MIR tasks and six benchmark datasets. By formalising each representation within a unified experimental framework, we isolate the contribution of biography sources and assess their interaction with acoustic features.

Our results show that aggregated artist-level textual augmentations provide complementary semantic information that consistently enhances multimodal representations across most evaluated settings. Improvements are particularly pronounced in genre classification and in larger-scale emotion recognition datasets, while gains in music tagging are primarily reflected in ranking-based metrics. These findings demonstrate that textual knowledge does not replace audio representations, but instead augments them with high-level contextual information.

Through targeted ablation studies, we further demonstrate that the observed improvements cannot be attributed solely to stronger audio encoders, generic metadata, or explicit genre mentions in biographies. Notably, the role of explicit lexical cues is dataset-dependent: on GTZAN, masking genre terms reduces the multimodal gain substantially, whereas on FMA-Small the drop is minimal. This suggests that biographies contribute through a mixture of surface-level lexical overlap and deeper contextual semantics, with the balance shifting toward the latter in larger-scale settings.

Having established the effectiveness of combining audio and textual representations, the following chapter investigates whether such knowledge can be internalised through pretraining strategies, enabling improved performance even in settings where textual augmentation is unavailable at inference time.

Chapter 6

Pretraining audio-text models

In Chapter 5, we show that artist biographies improve performance when combined with audio representations. While this multimodal fusion approach confirms the value of artist knowledge, it also requires access to biographies during inference, limiting its applicability in settings where such metadata is unavailable, incomplete, or unreliable.

This chapter investigates whether artist-level textual knowledge can instead be internalised through generative pretraining. Rather than supplying biographies as auxiliary inputs during inference, we ask whether an audio-text model can learn semantically enriched internal representations by being pretrained to reconstruct structured artist-level text from audio. To study this question, we extend the Whisper architecture introduced by Radford et al. [41] and adapt it to a music-specific pretraining regime in which acoustic inputs are paired with artist-level textual representations.

This objective differs from conventional supervised pretraining based on curated categorical labels. Artist biographies are semantically dense and typically encode multiple dimensions of musical identity, including genre affiliations, stylistic characteristics, historical context, and collaborations. Learning to predict such structured textual descriptions from audio, therefore, requires the model to capture higher-level semantic attributes beyond those needed for narrow discriminative tasks.

We compare direct fine-tuning from a speech-pretrained baseline with fine-tuning after an additional stage of biography-driven pretraining, on genre and emotion classification benchmarks. This is an exploratory study rather than a finished pretraining recipe: Chapter 5 established that biographical knowledge helps when it is supplied to the model, and the question here is whether the same knowledge can be absorbed instead, so that it still shapes the model's behaviour once the text is taken away. The two configurations differ in a single controlled variable, the presence of the additional pretraining stage, but the

evidence they can produce is directional rather than conclusive, for reasons we set out alongside the results in Sections 6.4.1 and 6.4.3.

6.1 Task reformulation

The pretraining objective adopted in this chapter frames music understanding as a conditional text generation problem, in which the model learns to reconstruct structured textual representations (e.g., artist biographies and metadata) from audio. To maintain methodological consistency between pretraining and downstream evaluation, the tasks considered in this chapter are reformulated within the same generative framework.

Rather than treating genre and emotion prediction as conventional classification problems with task-specific output heads, we recast them as conditional text generation tasks. Given an audio input and a task-specific textual prompt, the model is trained to generate the corresponding label as text. This unified formulation ensures that both pretraining and fine-tuning operate under the same autoregressive objective, avoiding architectural modifications between stages and enabling performance differences to be attributed to the pretraining regime.

This reformulation does not alter the semantic content of the prediction tasks, but only their output representation. Discrete class labels are expressed as textual tokens, and evaluation is performed by mapping generated outputs back to their corresponding categories. Importantly, this design is not intended to improve performance through prompt engineering, but to preserve consistency between pretraining and fine-tuning.

The same datasets introduced in Chapter 3 for genre classification and emotion recognition are used in this chapter, with the only modification being the reformulation of outputs into textual form.

6.1.1 Evaluation protocol

For each task, the model receives an audio input paired with a short textual prompt specifying the prediction objective (e.g., "Predict the genre of this track:"). During fine-tuning, the model is trained using teacher forcing to generate the ground-truth textual label. At inference time, generation is constrained to a short sequence corresponding to the predefined label space.

To compute evaluation metrics, generated outputs are mapped to valid labels within the task vocabulary, and standard classification metrics are computed on the resulting predictions.

6.1.2 Genre classification

Genre classification is reformulated as a single-label text generation task. For each audio input, the model is prompted to generate a single genre label from the

dataset's predefined set. The target output corresponds exactly to the textual representation of the ground-truth class (e.g., "rock", "jazz", "hip-hop").

Since genre labels are inherently categorical, this reformulation preserves the underlying task structure while modifying only the output modality. Performance is evaluated using accuracy over the discrete label set, as defined in Chapter 3.

6.1.3 Emotion classification

As described in Section 3.3, emotion recognition datasets represent affect along continuous valence and arousal dimensions. However, autoregressive text generation operates over discrete vocabularies, requiring categorical targets. To ensure compatibility with the generative framework adopted in this chapter, continuous emotion annotations are discretised into a finite set of categories.

Specifically, each point in the valence–arousal space is mapped to one of four affective classes: sad, happy, tender, or angry. The mapping strategy is illustrated in Figure 6.1. The quadrant boundaries are placed at the empirical mean of each dimension within each dataset rather than at the nominal midpoint of the annotation scale, for the reasons set out in Section 6.3.1, where we also report the resulting class distributions. This transformation defines a derived classification task that preserves the underlying affective structure while enabling prediction through textual label generation.

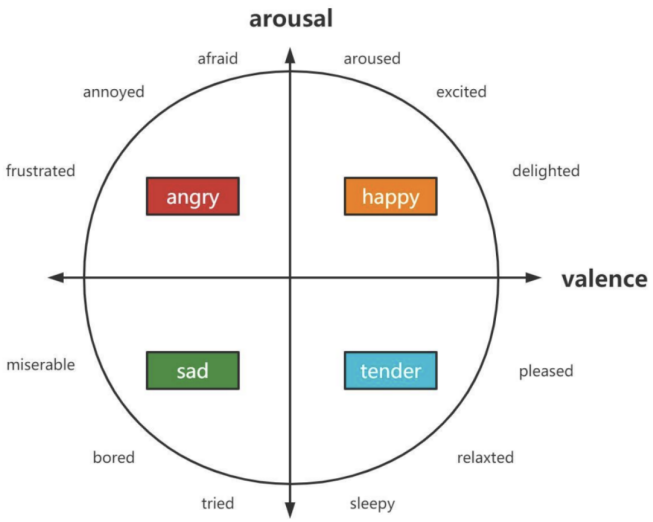


Figure 6.1: Valence–arousal plane annotated with the four discrete emotion categories (angry, happy, sad, tender) used for the classification task. Adapted from Russell [44].

This reformulation allows emotion recognition to be treated analogously to genre classification: the model is prompted to generate a textual label corresponding to the discretised emotion category. Although this introduces a discretisation step, it preserves the core objective of predicting affective characteristics from audio. Performance is evaluated using accuracy, since the task is cast as single-label classification.

6.1.4 Scope relative to the benchmark suite

The reformulations above imply two deliberate deviations from the evaluation protocol defined in Chapter 3, which we make explicit here.

Emotion recognition is evaluated using accuracy rather than R^2. The R^2 metric presupposes continuous-valued predictions, whereas an autoregressive decoder emits discrete textual tokens; once valence–arousal annotations are discretised into the four circumplex quadrants, accuracy over those categories is the natural metric. As a consequence, the emotion results in this chapter are not directly comparable to the R^2 scores reported in Chapter 5, and we instead report the VGGish baseline under the same discretised formulation to preserve internal comparability.

Continuous-valued outputs can be coaxed from a generative model, for instance by having it emit numerals and parsing them back to a scalar, or by reading a scalar off the decoder's token probabilities, and such a variant would restore formal comparability with the R^2 figures in Chapter 5. We chose not to pursue this for conceptual rather than practical reasons. Valence and arousal are not scalar quantities that happen to be recorded on a bounded interval; they are positions on a semantic scale, and the numeric annotation is an instrument for eliciting a judgement rather than a measurement of a physical property. A model emitting "0.62" for valence is not performing the same operation as a regression head trained against a squared-error objective, but producing a token whose numeric interpretation is imposed externally and whose ordering relation to neighbouring tokens it has no reason to respect. Fitting such a procedure to R^2 would report a number that looks comparable to a regression score while resting on a different underlying operation, which is the kind of confound the framework in Chapter 3 was designed to avoid. Discretising into the four circumplex quadrants is lossier but more honest: it asks the model for the categorical judgement that generative decoding can express, and makes the loss of resolution explicit rather than concealing it inside a continuous-looking metric. The cost is that the emotion columns cannot be read against Chapter 5, and we quantify what the discretisation itself costs in Section 6.3.1.

Music tagging is not evaluated in this chapter. The reason the experiments were not reported is empirical: we were unable to obtain acceptable *baseline* performance on tagging under the generative reformulation. Before the effect of biography pretraining could be assessed, the WhisperExt baseline had

to reach a level at which the reformulated task was being performed at all, and it did not. Results were very poor in absolute terms, falling far below the audio-only baselines reported in Chapter 5 for the same datasets, and generation of tag sets was unstable, with the model tending towards short, high-frequency tag sequences that varied little across inputs. This is a plausible failure mode for the formulation: predicting a variable-size set requires the decoder to learn both which tags apply and how many, and with 50 and 56 tag vocabularies and strong label imbalance the objective provides little pressure to do either well.

We report this as a negative result rather than omitting it, since a baseline that does not work is itself evidence that the generative reformulation is a poor fit for multi-label tagging at this scale and with this architecture. A comparison built on a non-functional baseline would have been misleading whichever configuration won. We therefore restrict the evaluation to the single-label tasks, and return to the implications of this restricted scope in Section 6.4.3.

6.2 Experimental setup

6.2.1 Model architecture

All experiments in this chapter build upon a speech-pretrained Whisper encoder-decoder architecture [41]. Initialising all models from the same pretrained checkpoint ensures identical sequence modelling capacity and linguistic priors across configurations. The only difference between models is whether an additional stage of music-specific generative pretraining is applied before downstream fine-tuning.

Due to computational constraints, we use the `whisper-small` variant of the publicly available Whisper model as the base checkpoint for all experiments.

Why Whisper, and when. The experiments in this chapter were designed and run in **spring and summer 2023**, and the choice of backbone should be read against the model landscape available at that time rather than the present one. This caveat is not incidental: audio and music representation learning moved quickly in the period immediately after these experiments, and several models that would now be obvious candidates either did not exist or had no released weights when the work was carried out. We state the date explicitly so that the choice can be assessed on the information available when it was made.

The research question narrows the field considerably. Because the pretraining objective is conditional text generation from audio (Section 6.1), it requires a pretrained decoder as well as a pretrained audio encoder, not merely a pretrained representation.

The principal alternative available to us was wav2vec 2.0 [3], which we considered and rejected for two reasons. Architecturally, wav2vec 2.0 is an encoder-only model, as are AudioSet-pretrained classification encoders such as the VGGish model used in Chapter 5; adopting one would have required training a text decoder from scratch, reintroducing a large randomly initialised component into

an experiment whose design depends on the two configurations differing only in the presence of the biography-pretraining stage. Empirically, we did not find wav2vec 2.0 competitive with the Whisper encoder for our purposes, which made the additional cost of building a decoder around it difficult to justify. Beyond wav2vec 2.0 the field was thin, as most of the pretrained audio and music models that would now be natural candidates had not yet been released.

Whisper satisfied the remaining requirements. It is an encoder-decoder trained end-to-end on a very large and deliberately heterogeneous audio collection, so both halves of the architecture arrive pretrained and mutually adapted. Its front-end is a log-mel spectrogram, matching the acoustic representation used throughout this thesis (Section 3.1) and keeping the input pipeline consistent across chapters. Its training regime prioritised robustness to varied and noisy audio over clean read speech, making it a more plausible starting point for musical audio than models trained on curated speech corpora. Weights were also publicly released at several sizes, which supports reproducibility and allowed us to work within our compute budget using the `small` variant.

The obvious objection is that Whisper is a speech model and music is not speech. We accept this. The acoustic prior it brings is imperfectly matched to the domain, and a comparable encoder-decoder pretrained on music would have been preferable, but no such model with released weights existed when this work was carried out; music-specific representation models such as MERT [29] appeared subsequently. Pretraining one ourselves was well outside the compute available for an exploratory study. We mitigate the mismatch in two ways. The lyrics representation is excluded from the pretraining targets, so the objective does not simply reinforce the speech-transcription behaviour already present in the checkpoint (Section 6.2.2), and both configurations start from the identical checkpoint, so whatever the speech prior contributes it contributes equally to both and cannot account for the difference between them. What the mismatch does bound is the absolute performance in Table 6.1; whether these results would transfer to a music-pretrained backbone remains open, and we return to it in Section 6.4.3.

Figure 6.2 illustrates the model architecture used in this chapter. We now describe each component in detail.

Audio encoder

The audio encoder corresponds to the Whisper encoder. Input audio is first converted into log-mel spectrogram representations and processed by an initial convolutional projection layer, followed by a stack of Transformer encoder blocks with multi-head self-attention and feed-forward (MLP) layers. The encoder outputs a sequence of contextualised audio embeddings.

Text encoder

To enable explicit conditioning on textual prompts, we introduce an additional text encoder. This encoder is initialised from a pretrained mBART model,

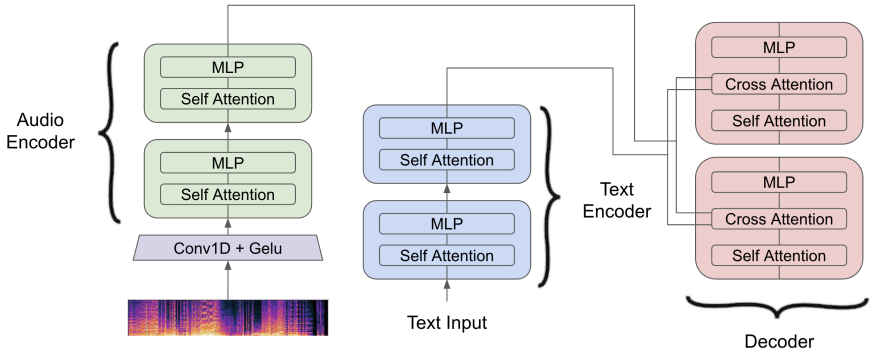


Figure 6.2: Extended Whisper architecture used to evaluate the impact of music-specific pretraining on genre and emotion classification tasks.

providing strong sequence modelling priors for textual inputs. The text encoder processes prompt text and produces contextualised textual embeddings, which are aligned with the audio representations at the encoder level. This design allows task prompts to influence the representation before decoding, rather than relying solely on decoder-side conditioning.

Multimodal fusion and decoder

The outputs of the audio encoder and text encoder are concatenated along the sequence dimension to form a unified multimodal representation. This combined embedding sequence is provided to the shared Transformer decoder via cross-attention. The decoder, inherited from Whisper, performs autoregressive generation conditioned on the multimodal representation.

This design enables task conditioning through the encoder while preserving a single shared generative decoder across both pretraining and fine-tuning. All model parameters, including those inherited from Whisper and mBART, remain trainable throughout both stages.

6.2.2 Pretraining procedure

We select the Million Song Dataset (MSD) [4] as the pretraining corpus. With $993,175$ samples, it provides a large and diverse collection of tracks for generative supervision. The dataset is annotated using the text-combined representations introduced in Chapter 5, pairing each audio track with a structured textual representation that concatenates track-level metadata and artist biographies.

The lyrics representation is deliberately excluded from this pretraining stage. Since Whisper is originally pretrained for speech recognition, incorporating lyrics (which are linguistically aligned with speech transcription) could bias the model toward reinforcing speech-to-text associations rather than encouraging

the acquisition of higher-level musical knowledge. By restricting supervision to metadata and biography text, the pretraining objective focuses on semantically enriched, music-specific information rather than transcription-like signals.

While explicit artist identifiers are not provided to the model, this structure may encourage the emergence of artist-discriminative acoustic representations. Due to the corpus construction, there is partial overlap between the artists used during pretraining and those appearing in downstream evaluation datasets. We will discuss in-depth the impact of this overlap in Section 6.4.1. The objective of this study is therefore not to assess strict cross-artist generalisation, but to evaluate whether semantically enriched generative supervision improves fine-tuning performance.

Pretraining was performed for $200,000$ optimisation steps, corresponding to approximately 0.2 epochs over the MSD corpus.

Pretraining objective

Pretraining follows the span-masked conditional generation objective introduced in [42]. In this formulation, contiguous spans of tokens in the target text are replaced with sentinel mask tokens, and the model is trained to reconstruct the masked spans.

In our setting, the target text corresponds to the structured artist-level representation associated with each track (i.e., metadata and biography text). The model receives the corrupted text and the audio representation as input and is trained to predict the masked spans conditioned on both modalities.

This objective encourages the model to learn contextualised textual representations while leveraging acoustic information to guide reconstruction. Compared to standard autoregressive generation, span-masked prediction reduces the risk of trivial memorisation and promotes tighter alignment between audio-derived features and high-level textual semantics. The model is trained end-to-end using a cross-entropy loss over the masked spans.

Figure 6.3 illustrates an example of a masked pretraining instance. The original text is corrupted by masking contiguous spans, and the model is trained to reconstruct the missing content.

Since the pretraining targets include artist-level descriptions and metadata, the model may learn to associate acoustic patterns with high-level semantic attributes such as style, genre, or artist identity. Notably, masked spans may include explicit artist identifiers, potentially providing a signal linking audio characteristics to specific artists. We discuss the implications of this effect in Section 6.4.1.

Unlike standard text-only span-masked pretraining, reconstruction in this setting is conditioned on audio features, requiring the model to use acoustic cues to resolve masked semantic content.

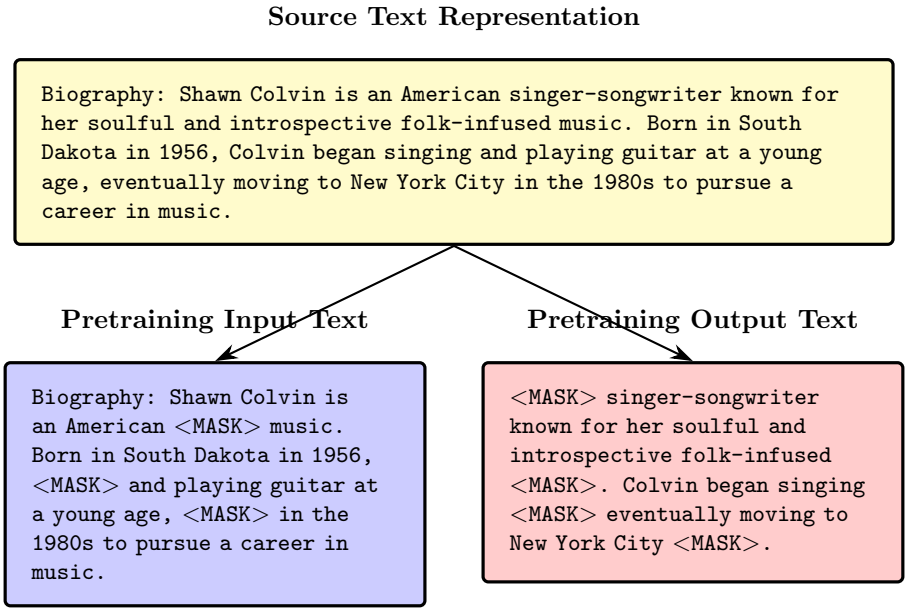


Figure 6.3: Span-masked pretraining objective. The source text is corrupted by masking spans to form the input sequence (Pretraining Input Text), and the model is trained to reconstruct the masked spans (Pretraining Output Text).

6.2.3 Fine-tuning procedure

Following pretraining, the model is fine-tuned on downstream genre and emotion classification tasks reformulated as conditional text-generation problems, as described in Section 6.1.

Two configurations are compared:

- **Direct Fine-Tuning:** Speech-pretrained Whisper model fine-tuned directly on downstream tasks.
- **Span-Mask-Pretrained Fine-Tuning:** Speech-pretrained Whisper model first pretrained on span-masked pretraining, then fine-tuned on downstream tasks.

Both configurations share identical architecture, optimisation settings, and evaluation protocol. All parameters remain trainable during fine-tuning, ensuring that any observed performance differences can be attributed to the additional pretraining stage rather than architectural changes.

6.3 Results

Table 6.1 summarises the performance of the evaluated configurations across genre and emotion classification benchmarks. Results are reported for VGGish

as a reference baseline, alongside Whisper-based models, to contextualise performance relative to the representation-based approaches studied in previous chapters.

	GTZAN	**FMA-Small**	**EmoMusic**	**Deezer**
	Accuracy	Accuracy	Accuracy	Accuracy
VGGish	77.2	52.7	63.2	49.3
WhisperExt	79.0	52.7	62.4	52.1
WhisperExt + SMP	**85.1**	**57.5**	**66.4**	**53.7**

Table 6.1: Experimental results of the impact of music knowledge pretraining on the performance of genre and emotion classification.

We first compare the baseline Whisper configuration (WhisperExt) with VGGish. Overall, both models exhibit comparable performance, with WhisperExt yielding a modest average relative improvement of +1.7% across datasets. This suggests that the generative text-to-text formulation does not inherently degrade performance, despite introducing a more complex prediction mechanism. However, differences emerge across tasks. On EmoMusic, WhisperExt performs slightly worse than VGGish (−1.3% relative), indicating that the generative formulation may be more sensitive in low-data regimes. In contrast, on Deezer, which provides a substantially larger training set, WhisperExt achieves a relative improvement of +5.7%, suggesting that the architecture benefits from increased data availability.

We now analyse the impact of biography-driven pretraining. Incorporating this additional pretraining stage (WhisperExt + SMP) yields consistent improvements across all evaluated datasets, with an average relative gain of +6.6% over direct fine-tuning. The fact that gains are observed in all settings indicates that the effect is robust and not tied to a specific dataset or task configuration.

The most pronounced improvements are observed in genre classification, particularly on GTZAN. This behaviour is consistent with the hypothesis that artist-level textual supervision encodes high-level stylistic information (e.g., genre and musical identity) that transfers effectively to downstream tasks. It also aligns with observations from Section 5.5, where models trained with large-scale pretraining were seen to hold their largest advantage over VGGish in the lowest-resource setting.

In contrast, Deezer exhibits the smallest relative gain from biography pretraining (+3.1%). While still positive, several factors may explain this more moderate improvement. First, the discretisation of emotion from continuous valence–arousal annotations introduces an additional abstraction layer, potentially weakening alignment with biography-derived semantic information. Second, affective characteristics are less directly encoded at the artist level than stylistic attributes such as genre, limiting the transferability of biography-based supervision. More generally, the task-agnostic nature of the pretraining objective may favour the learning of stable, artist-level properties over track-specific

emotional variation.

Overall, these results provide consistent evidence that biography-driven generative pretraining induces beneficial inductive biases for downstream music classification. While the magnitude of improvements varies across tasks and datasets, the uniform direction of the effect suggests that long-form, semantically rich textual supervision contributes meaningfully to representation learning.

Due to computational constraints, each configuration was trained only once; therefore, the reported results do not account for variance across random initialisations and should be interpreted as indicative rather than definitive.

These results are broadly consistent with the hypothesis that biography-driven pretraining yields improvements over direct fine-tuning, but that these gains are smaller and less consistent than those obtained through explicit multimodal augmentation in Chapter 5. The average relative improvement from pretraining (+6.6%) is modest compared to the multimodal fusion gains reported earlier, consistent with the expectation that implicit internalisation requires the model to compress contextual knowledge into its representations rather than accessing it directly.

6.3.1 Distribution of music across the circumplex quadrants

The emotion results above depend on a discretisation step, and that step is not neutral. How the tracks are distributed across the four quadrants determines both what the accuracy figures mean and how much can be learned about each class, so we examine that distribution here.

Where the decision boundaries lie. A point in the valence–arousal plane must be assigned to a quadrant relative to some origin. Both datasets are annotated on scales whose nominal midpoint does not coincide with the empirical centre of the data: Deezer's annotations are approximately zero-centred but skewed towards positive arousal (mean arousal 0.19), and EmoMusic's nine-point scale has an empirical mean of 5.01 for valence and 4.79 for arousal rather than the nominal 5.0 for both. Using the nominal midpoint would therefore introduce an artificial bias towards the high-arousal classes. We instead place the origin at the empirical mean of each dimension within each dataset, so that the quadrants partition the observed distribution rather than the nominal scale. This is a defensible choice, but it also means the class boundaries are dataset-specific and the resulting labels are not directly comparable across the two datasets.

The resulting distribution. Table 6.2 reports the number of tracks in each quadrant. Neither dataset is balanced, and the imbalance is structured rather than incidental. In both, the dominant classes are the congruent ones, high arousal with high valence (happy) and low arousal with low valence (sad), which

together account for 64.3% of Deezer and 76.1% of EmoMusic. The incongruent combinations, high-arousal/low-valence (angry) and low-arousal/high-valence (tender), are correspondingly rare. This reflects a known property of affective annotation: valence and arousal are not perceived independently, and their empirical correlation is positive in both datasets ($r = 0.25$ for Deezer, $r = 0.60$ for EmoMusic). Music that is simultaneously calm and cheerful, or agitated and bleak, is genuinely less common than the diagonal cases, and annotators also tend to rate the two dimensions consistently. The circumplex model treats valence and arousal as orthogonal axes, but the data populating it are not distributed orthogonally.

The imbalance is far more severe for EmoMusic, which is also the smaller dataset by an order of magnitude. Its two minority classes contain 96 and 82 tracks in total, 12.9% and 11.0% of the dataset, so of the 619 training samples only around eighty are labelled "angry" and seventy "tender". Any conclusion about those classes rests on a thin basis, and the EmoMusic column of Table 6.1 is in practice dominated by the two diagonal classes.

Quadrant	**Label**	**Deezer**		**EmoMusic**	
Q3: low arousal, low valence	sad	5,996	32.7%	284	38.2%
Q1: high arousal, high valence	happy	5,793	31.6%	282	37.9%
Q2: low arousal, high valence	tender	3,718	20.3%	82	11.0%
Q4: high arousal, low valence	angry	2,821	15.4%	96	12.9%
Total		**18,328**		**744**	

Table 6.2: Distribution of tracks across the four circumplex quadrants after discretisation, using the empirical mean of each dimension as the origin. The Deezer total covers the tracks available to this chapter's experiments, and is therefore slightly lower than the 18,644 listed in Chapter 3. Both datasets are dominated by the congruent (happy/sad) diagonal, and EmoMusic is both smaller and more skewed than Deezer.

Boundary effects. Discretisation also imposes hard categories on a continuous space, and does so most arbitrarily where the data are densest. Because both distributions are unimodal and centred near the origin, a substantial fraction of tracks lie close to a decision boundary: 6.3% of Deezer tracks and 14.5% of EmoMusic tracks fall within one tenth of a standard deviation of one of the two axes, rising to 21.6% and 31.2% respectively within a quarter of a standard deviation. For these tracks the assigned label is determined by annotation differences far smaller than the disagreement typically observed between annotators, and two perceptually indistinguishable tracks can receive different targets. This produces a floor of irreducible label noise that is concentrated in the centre of the space and is larger for EmoMusic, whose annotations are averaged over the full duration of each excerpt (Section 3.3) and are therefore already smoothed towards the centre.

Implications for the reported accuracies. These properties mean the accuracy figures in Table 6.1 should not be read against a 25% chance level. The appropriate reference points are the majority-class predictors, which achieve 32.7% on Deezer and 38.2% on EmoMusic. All configurations exceed them by a clear margin, with the best model reaching 53.7% and 66.4% respectively, so the models are learning genuine affective structure rather than exploiting the class prior. The absolute values should nonetheless be treated with care. Accuracy under this imbalance is dominated by the two majority classes, so a model could improve on the metric while remaining poor on the minority quadrants, which the aggregate figure conceals. The centre-concentrated label noise places a ceiling on attainable accuracy independent of the model. And because the datasets have different class distributions and dataset-specific boundaries, their accuracy columns are comparable across configurations within a column but not with each other. None of this undermines the comparison the chapter makes, which holds the discretisation fixed and varies only the pretraining stage.

6.3.2 The complexity of the reformulated tasks

The gains reported above are consistent in direction but moderate in size, which reflects how much harder these tasks are than their counterparts in Chapter 5.

The central difficulty follows directly from the research question. In Chapter 5 the biography is an input: the model is told who the artist is and what has been written about them, and its task is to combine that description with the acoustic signal. Here it is given neither. At inference time the model receives only the waveform and a task prompt, with no artist name, no artist identifier, and no biography, so any biographical knowledge it brings to bear must have been compressed into its parameters during pretraining and recovered from acoustic evidence alone. To exploit the knowledge that a particular artist works in a given stylistic tradition, the model must first recognise from the audio that the track plausibly belongs to that artist, or at least to a region of acoustic space populated by similar artists, and then apply what it absorbed about that region. It must learn this audio-to-artist association as a by-product of the reconstruction objective, without ever being supervised on artist identity and without a textual channel to fall back on when the acoustics are ambiguous. Chapter 5 required the model to use a description; this chapter requires it to have internalised one and to know when it applies. The smaller effect sizes are consistent with that additional difficulty, and the comparison against the multimodal gains should be read with the asymmetry in mind.

The emotion benchmarks add a second difficulty, since their targets are less well defined than a genre label. As Section 6.3.1 shows, discretising a continuous, unimodally distributed affective space places class boundaries through the densest part of the data, so a non-trivial fraction of examples carry labels that could plausibly have been otherwise. On EmoMusic this compounds with the temporal aggregation described in Section 3.3: the target is the mean of a valence–arousal trajectory over the whole excerpt, so a track whose affect changes over its duration is represented by an average that may correspond to

no moment in the music. Part of the residual error on these tasks is therefore attributable to the construction of the target rather than to the representation being evaluated.

6.4 Limitations and future work

While the results presented in this chapter provide encouraging evidence for biography-driven generative pretraining, several limitations constrain the scope of the conclusions that can be drawn. These limitations primarily concern dataset construction, the nature and structure of textual supervision, and the modelling assumptions introduced by the generative reformulation. Addressing these aspects constitutes a natural direction for future research.

6.4.1 Dataset overlap and artist supervision

Pretraining was conducted on the subset of the Million Song Dataset (MSD) for which textual metadata and biography information were available. As a result, varying degrees of artist-level overlap exist between the pretraining corpus and downstream benchmarks. Table 6.3 reports the percentage of samples in each downstream dataset that have an artist that also appears in the MSD-based pretraining corpus.

GTZAN	**FMA-Small**	**EmoMusic**	**Deezer**
87.0%	6.5%	8.1%	100.0%

Table 6.3: Percentage of evaluation samples whose associated artist appears in the MSD pretraining corpus.

The degree of overlap varies substantially across datasets, ranging from minimal overlap in FMA-Small and EmoMusic to complete overlap in Deezer. This complete overlap in the case of Deezer is expected, as the dataset is derived from the MSD dataset. This distribution enables a partial analysis of the relationship between artist familiarity and downstream performance gains.

Interestingly, improvements from biography-driven pretraining are observed both in high-overlap settings (e.g., GTZAN) and in low-overlap settings (e.g., FMA-Small and EmoMusic). While artist-level familiarity may contribute to gains in highly overlapping datasets, the presence of improvements in low-overlap regimes suggests that the observed benefits cannot be attributed solely to memorisation effects. Nevertheless, the results should be interpreted as reflecting improved representation learning under partially shared distributions rather than strict generalisation to entirely unseen artists.

Future work should enforce fully disjoint artist splits during pretraining and evaluation to isolate cross-artist transfer effects more rigorously.

6.4.2 Sources of artist biographies

The pretraining objective relies on the text-combined representation introduced in Chapter 5, which aggregates multiple textual sources, including metadata and artist biographies. Among these sources, LLM-generated biographies constitute a significant component, providing structured, consistently formatted descriptions for artists.

While this design enables scalable and uniform supervision, it introduces some limitations. The use of a fixed aggregation strategy constrains the diversity and structure of the textual supervision. Although the combined representation integrates multiple sources, it imposes a specific formatting and concatenation scheme that may influence how information is prioritised during pretraining. This raises the question of whether the observed improvements stem from the underlying knowledge content or from the representation's particular structure.

Moreover, this work does not disentangle the contribution of individual textual sources within the combined representation. In Chapter 5, different biography sources were shown to provide complementary information; however, their respective roles in the pretraining setting remain unexplored.

Future work should investigate alternative and more diverse sources of supervision, such as critical reviews, interviews, production notes, or multilingual corpora. In addition, systematically varying the composition and structure of textual representations would help determine whether performance gains stem from specific knowledge sources or from broader semantic enrichment.

6.4.3 Generative reformulation limitations

Recasting downstream tasks as conditional text generation ensures architectural consistency between pretraining and fine-tuning, allowing the effect of generative supervision to be isolated. However, this modelling assumption introduces constraints, and it is the reason the evaluation in this chapter covers only part of the benchmark suite established in Chapter 3 (Section 6.1.4). Continuous regression targets must be discretised, replacing R^2 with accuracy for the emotion task; and multi-label tagging is excluded because we could not obtain a working baseline for tagging under this formulation at all.

The evidence presented here supports biography-driven pretraining for *single-label semantic tasks*, but cannot draw any conclusion on multi-label tagging. We cannot exclude that a different prompt design might be able to perform music tagging, but this is not the focus of the chapter.

Future work should investigate hybrid architectures that retain biography-driven pretraining while downstream prediction is performed via task-specific classification heads. More broadly, it remains an open question where within the model the injected musical knowledge exerts the greatest influence, and whether explicit language modelling is necessary to preserve the inductive bias introduced during biography pretraining.

6.4.4 Scale of pretraining

All experiments in this chapter build on a Whisper backbone pretrained on large-scale speech transcription data, followed by additional biography-driven supervision. While this additional pretraining demonstrates consistent improvements, its scale remains modest relative to contemporary large-scale multimodal pretraining regimes.

Exploring the effects of scaling the quantity and diversity of external textual supervision may reveal stronger transfer effects or emergent behaviours. A systematic investigation of scaling dynamics would clarify whether biography-driven generative supervision follows the same scaling laws observed in other multimodal foundation models.

6.5 Contributions

This chapter makes the following contributions:

- **Span-masked multimodal generative pretraining:** We introduce a pretraining strategy that leverages structured textual representations to internalise musical knowledge within an audio-text model, extending beyond the feature-level multimodal fusion explored in Chapter 5.
- **Unified generative task formulation:** We propose a consistent text-to-text reformulation of genre and emotion classification, aligning pretraining and fine-tuning under a shared conditional generation objective without requiring task-specific architectural modifications.
- **Empirical evidence of semantic transfer:** We demonstrate that biography-driven span-masked pretraining yields consistent performance improvements across multiple datasets, indicating that long-form textual supervision induces beneficial inductive biases for downstream music classification.

6.6 Summary

This chapter investigates whether musical knowledge, previously introduced as augmented textual representations, can be internalised through generative pretraining rather than provided explicitly at inference time. Building on a speech-pretrained Whisper architecture, we extend the model with a text encoder and train it to reconstruct structured textual representations from audio using a span-masked objective. This formulation allows us to assess whether semantically enriched supervision induces transferable inductive biases in audio representations.

Across genre and emotion classification benchmarks, span-masked pretraining yields consistent improvements over direct fine-tuning from a speech-pretrained baseline, with an average relative gain of +6.6%. Gains are particularly pronounced in genre classification, indicating that artist-level textual supervision aligns well with stylistic attributes encoded in music. The consistency of the

effect across four datasets, spanning both high and low artist overlap with the pretraining corpus, suggests that long-form textual supervision provides a meaningful signal for downstream learning rather than an artefact of any single benchmark.

These conclusions remain exploratory, and several constraints bound them: a single run per configuration, artist overlap between pretraining and evaluation data, a speech-pretrained rather than music-pretrained backbone, and coverage of only the single-label tasks, since the generative reformulation did not yield a workable baseline for tagging. What the chapter establishes is that biography-driven supervision continues to influence a model after the text is removed, that the internalisation route is therefore viable, and that its effect is smaller than direct injection. This is a narrower claim than the multimodal results of Chapter 5 support, and it also identifies the conditions a properly powered study would need to control.

Taken together, these results nevertheless extend the thesis beyond multimodal fusion. Structured artist-level textual representations not only complement audio features when used as inputs, but also appear to shape the internal structure of learned representations when used as supervision. This indicates that external knowledge can be internalised within audio-text models, and identifies knowledge-guided pretraining as a direction worth pursuing at greater scale for developing semantically grounded systems in music information retrieval.

Chapter 7

Discussion and conclusion

This thesis investigates whether structured artist-level textual knowledge, in the form of biographies, can enhance music representations for semantic music information retrieval (MIR) tasks. Across a unified evaluation framework spanning genre classification, emotion recognition, and music tagging, we systematically evaluate the impact of incorporating artist biographies derived from multiple sources. To support this analysis, we introduce two reproducible corpora of artist biographies and assess their coverage and reliability.

Across controlled experimental settings, we find that artist biographies improve performance on most evaluated tasks when integrated with audio representations, with the notable exception of low-resource emotion recognition, where gains remain inconclusive. Furthermore, we provide evidence that such knowledge can be internalised through generative pretraining, yielding improvements even when textual information is unavailable at inference time.

More broadly, the results suggest that artist biographies encode structured semantic information that can influence model behaviour in two distinct ways: either through direct multimodal augmentation at inference time or by shaping inductive biases during pretraining. While these findings should be interpreted within the methodological and dataset constraints described in this thesis, the consistent patterns observed across tasks indicate that structured textual supervision constitutes a promising direction for knowledge-aware music representation learning.

The principal contribution. This thesis makes several contributions: two resources, an evaluation framework, an empirical study, and a conceptual distinction. The principal contribution is the **empirical demonstration that artist-biography augmentation measurably improves multimodal music representations**, established in Chapter 5 across three tasks and six benchmark datasets, and supported by the leakage-controlled ablations that separate genuine semantic contribution from surface-level lexical overlap.

The multimodal experimentation is the theme and the core of this work, and the remaining contributions all feed into it. The evaluation framework of

Chapter 3 exists to make the comparison interpretable, by holding the architecture and protocol fixed so that the representation is the only variable. The corpora of Chapter 4 exist because the comparison could not otherwise be run reproducibly: without an openly constructible source of artist-level text, the experiment would have depended on resources that cannot be shared. The pretraining study of Chapter 6 asks what happens to the same knowledge when it is absorbed rather than supplied, extending the central finding along a second axis.

The demonstration is *evidence towards*, not proof of, the relevance of contextual knowledge in MIR. The experiments show that a particular form of contextual knowledge, supplied in a particular way, improves performance on a particular set of semantic tasks under controlled conditions, and that through the masking ablations this improvement is not reducible to the target label appearing in the input text. They cannot show that contextual knowledge is necessary, that biographies are the best vehicle for it, or that the effect would survive on a substantially stronger acoustic encoder. Artist biographies are best understood as *one possible vehicle* for embedding field knowledge into music models, convenient because they are abundant, artist-level, and openly obtainable rather than because there is anything privileged about biography as a genre. The broader proposition the thesis supports, that musical attributes such as genre, mood, and lineage are partly culturally constituted rather than fully latent in the waveform, and that models therefore benefit from knowledge the signal does not contain, is a direction this evidence motivates rather than a result it establishes.

We will now summarise the contributions of this thesis.

7.1 Summary of contributions

7.1.1 Evaluation methodology

This thesis first establishes a controlled and task-driven evaluation framework for assessing music representations in music information retrieval (MIR). Rather than introducing new prediction tasks, it consolidates a diverse set of established benchmarks across genre classification, emotion recognition, and music tagging into a unified pipeline. By fixing the model architecture, dataset splits, and evaluation metrics, the framework isolates the contribution of the representation itself, reducing confounding effects from architectural variation. This design enables systematic comparison of audio-only, textual, and multimodal representations under consistent experimental conditions.

This structured evaluation protocol underpins all subsequent experiments and provides a reproducible foundation for studying representation-level impact in MIR.

7.1.2 Building the artist biography corpora

To enable structured textual augmentation of music representations, this thesis introduces two reproducible corpora of artist-level biographies, constructed from complementary sources: community-authored web content and large language model-generated text. These corpora are designed to balance coverage, stylistic consistency, and accessibility, addressing limitations of existing encyclopedic resources that are either restricted by copyright or biased toward highly popular artists. The resulting datasets provide high coverage across multiple MIR benchmarks while remaining reproducible for research use.

Beyond corpus construction, the thesis proposes a factual reliability analysis grounded in Wikidata-based question answering. This evaluation quantifies the extent to which different biography sources encode verifiable biographical attributes and highlights systematic trade-offs between community-authored variability and LLM-generated consistency. Rather than assuming textual augmentation to be inherently beneficial, this analysis characterises the properties, biases, and limitations of each knowledge source, enabling informed interpretation of downstream experimental results.

7.1.3 Multimodal music representations

Building on the evaluation framework and textual corpora, the thesis systematically compares audio-only, metadata-based, biography-augmented, and aggregated textual representations across six benchmark datasets. The results demonstrate that artist-level biographies provide complementary semantic information to acoustic features, particularly in genre classification and large-scale emotion recognition settings. When combined with audio in a controlled multimodal architecture, aggregated textual representations yield improvements on five of six evaluated datasets, with gains on EmoMusic remaining inconclusive due to its limited training data.

Importantly, ablation analyses clarify that these improvements cannot be attributed solely to trivial metadata cues or architectural capacity. These findings support the conclusion that artist biographies encode high-level stylistic and contextual attributes that augment acoustic representations in a meaningful and measurable way.

7.1.4 Pretraining on artist biographies

Extending beyond multimodal fusion at inference time, this thesis investigates whether artist-level textual knowledge can be internalised through generative pretraining. By reformulating downstream tasks within a conditional text generation framework and pretraining a speech-initialised audio-text model to reconstruct textual representations from audio, the work explores whether structured textual supervision induces beneficial inductive biases. Experimental results show consistent improvements over direct fine-tuning on genre and emotion

benchmarks, suggesting that biography-driven pretraining influences learned representations even when textual input is unavailable at inference.

While the scale and dataset overlap constraints limit the scope of definitive conclusions, the observed pattern of improvements across tasks indicates that long-form semantic supervision constitutes a viable alternative to feature-level augmentation. Together with the multimodal experiments, these results highlight two complementary mechanisms for incorporating structured knowledge into MIR systems: explicit injection via multimodal integration and implicit learning via generative pretraining.

Taken together, these contributions establish structured artist-level textual knowledge as a practical and effective mechanism for enhancing music representation learning under controlled and reproducible conditions.

7.2 Discussion

7.2.1 Revisiting the research hypotheses

Chapter 1 introduced three research questions alongside explicit hypotheses about expected outcomes. These questions were made answerable by the controlled evaluation framework defined in Chapter 3, itself a methodological contribution rather than an open question: it enabled comparison of representations across genre classification, emotion recognition, and music tagging while holding the downstream architecture fixed, and it served that purpose throughout the thesis. Fixing the architecture does not eliminate all confounds: differences in coverage, alignment, and leakage risk across representations remain important interpretive considerations, and they recur below. Having presented the empirical findings in Chapters 4–6, we now revisit each hypothesis and assess the extent to which it was supported, complicated, or left unresolved by the evidence.

- **RQ1: Scalability and reliability of artist biographies.** The hypothesis that biographies could be scalably constructed was confirmed: the web-scraped corpus spans 18,912 artists, of which 8,484 carry biography text, and the LLM-generated corpus achieves over 90% coverage across all evaluation datasets. However, scalability comes with trade-offs that the hypothesis did not fully anticipate. LLM-generated biographies are factually denser and more consistent, but susceptible to entity ambiguity and biased toward well-documented artists. Web-scraped biographies offer greater diversity but are sparser and noisier. Neither corpus is uniformly reliable, and the factual evaluation in Chapter 4 highlighted that high coverage does not guarantee high-quality supervision.
- **RQ2: Multimodal improvements from biography augmentation.** The hypothesis that biographies would improve performance through complementary semantic cues was broadly confirmed, with multimodal fusion yielding gains across most tasks and datasets. As predicted, the improvements were strongest for genre classification and less consistent for affec-

tive tasks. However, the genre leakage analysis complicated the picture: on GTZAN, where 59.4% of biographies contain explicit genre terms, a substantial portion of the gain is attributable to direct lexical cues rather than deeper contextual information. On FMA-Small, the drop from masking genre terms was much smaller, suggesting that broader contextual semantics contribute more in larger-scale settings. The hypothesis was therefore partially confirmed: biographies help, but through a mixture of mechanisms whose relative contributions are dataset-dependent.

- **RQ3: Internalisation through pretraining.** The hypothesis that pretraining gains would be real but smaller than inference-time augmentation was directionally confirmed. Biography-driven pretraining produced consistent improvements over direct fine-tuning (+6.6% average relative gain), but these were indeed more modest than the multimodal fusion gains in Chapter 5. However, this finding remains preliminary: the chapter relies on a single training run with known artist overlap between pretraining and evaluation data, and no statistical testing was performed. The evidence is consistent with the hypothesis but not yet strong enough to be considered definitive.

7.2.2 Positioning within multimodal MIR research

Multimodal learning has become an important direction within Music Information Retrieval, with recent progress largely driven by large-scale audio-text pretraining and contrastive alignment strategies. These approaches typically leverage substantial quantities of weakly annotated or proprietary data to learn joint representations across modalities.

Within this landscape, this thesis examines the use of structured artist-level biographies as an additional source of semantic supervision. Rather than focusing primarily on scale, the work explores whether the semantic density and contextual coherence of biographical texts can augment music representations under controlled experimental settings. By incorporating artist-level information, the proposed framework introduces a hierarchical perspective in which relatively stable stylistic and contextual attributes inform track-level modelling.

The resulting approach provides a reproducible form of textual supervision accessible within academic settings. In this sense, the thesis contributes an alternative design axis for multimodal MIR research, centred on structured semantic knowledge rather than on dataset scale.

7.2.3 Artist biographies as textual representations

Mapping textual descriptions to music has long been a challenge in Music Information Retrieval, where textual supervision has typically relied on either lyrics or corpora curated by domain experts. Lyrics provide track-level, affective, and often noisy signals, whereas expert-crafted resources typically encode structured but task-specific annotations. Both approaches require substantial

effort to curate at scale, and large, high-quality corpora are frequently restricted or proprietary.

Within this context, this thesis evaluates artist biographies as an alternative source of textual supervision. Compared with traditional textual resources, biographies can be collected from community-authored sources or generated by large language models, making them more accessible for academic research. Empirically, they provide complementary semantic information, improving multimodal models under controlled experimental conditions.

A distinctive characteristic of biographies is the level of supervision they receive. Unlike lyrics, which operate at the track level and reflect instance-specific affective content, biographies describe artists at a broader scope. This introduces a form of hierarchical supervision: relatively stable stylistic and contextual attributes associated with an artist inform representations of individual tracks. As a result, biography-based representations tend to encode more persistent characteristics, such as genre affiliations, instrumentation, cultural influences, and temporal or geographical context, rather than transient properties of specific songs.

In this sense, biographies can be understood as a form of structured cultural discourse about music. They reflect how artists are situated within stylistic traditions, social contexts, and historical trajectories. Rather than directly describing the acoustic content of a specific track, they provide contextual framing that complements instance-level signals such as audio features or lyrics.

However, this representation paradigm also introduces limitations. Because supervision is anchored at the artist level, biography-based representations are less sensitive to intra-artist variability and may struggle to capture stylistic shifts or atypical releases. Additionally, the quality and richness of biographies are often correlated with an artist's popularity, introducing a potential bias against long-tail or emerging artists. These structural characteristics constrain the scope within which biography-based supervision can be expected to provide consistent benefits.

7.2.4 Injection vs learning of textual knowledge

A central distinction explored in this thesis concerns two mechanisms for incorporating artist-level textual knowledge into music representation learning: explicit biography injection at inference time and implicit knowledge internalisation through generative pretraining. Although both approaches yield performance improvements under controlled experimental settings, they operate through fundamentally different mechanisms and present distinct practical trade-offs.

Biography injection augments each music sample with artist-level textual information during inference. This approach requires maintaining a comprehensive, up-to-date collection of artist biographies, and its effectiveness depends on coverage. Consequently, unknown or newly emerging artists pose a challenge. At the same time, this design offers transparency and modularity: the textual input is explicit and controllable, and it can be updated independently of the

model parameters. The behaviour of the system can therefore be influenced by modifying or refining the textual resource itself.

In contrast, pretraining on biographies seeks to internalise structured contextual information into the model parameters through a span-masking objective. Rather than relying on textual input at inference time, this approach aims to shape the audio-text model's inductive biases during training. Empirically, the results suggest that biography-driven pretraining can transfer beneficial contextual patterns across tasks. However, this strategy reduces transparency and modularity: once knowledge is internalised, modifying model behaviour requires additional training, and the influence of specific textual signals becomes less directly interpretable.

Taken together, these findings illustrate two complementary paradigms for integrating structured knowledge into MIR systems. Explicit injection provides controllable, interpretable conditioning, whereas generative pretraining offers a pathway to embed contextual information into the representation itself. The relative suitability of each approach depends on deployment constraints, data availability, and the desired balance between controllability and representational autonomy.

7.2.5 Interpretation of empirical findings

Across the experiments presented in this thesis, a broad pattern emerges: artist biographies provide complementary semantic information that enhances performance on most evaluated semantic MIR tasks. The benefits are most pronounced for genre classification, where artist-level supervision aligns closely with stable stylistic attributes, and more modest for music tagging. Emotion recognition presents a mixed picture: gains are observed on the larger-scale Deezer dataset, but remain inconclusive on EmoMusic, where limited training data prevents reliable learning from textual signals. This distribution of gains reflects the level at which biographies operate, namely as relatively stable descriptors of artistic identity rather than track-specific signals.

Importantly, biographies do not replace acoustic information but complement it. Across tasks, multimodal combinations of audio and biography representations perform consistently as well as or better than unimodal baselines. This indicates that the textual and acoustic channels capture partially distinct aspects of musical semantics, with biographies contributing contextual and stylistic framing that augments instance-level acoustic detail.

Ablation studies further clarify the nature of this contribution. Masking explicit genre terms in biographies reduces but does not eliminate performance gains, indicating that improvements cannot be attributed solely to direct label leakage. Instead, the results suggest that biographies encode higher-level stylistic patterns and contextual associations that extend beyond explicit categorical cues.

Finally, the pretraining experiments reinforce this broader pattern. Although exploratory in scope, they indicate that biography-based supervision can influence learned representations even when textual input is unavailable

at inference time. This supports the interpretation that contextual knowledge derived from artist biographies can be internalised within multimodal models, shaping their inductive biases for genre and emotion inference.

Taken together, these findings characterise artist biographies as a form of structured contextual supervision that is particularly well aligned with stable, high-level semantic attributes in music representation learning.

7.3 Limitations

While the findings presented in this thesis provide evidence that artist biographies can enhance music representation learning under controlled settings, they should be interpreted within the scope and constraints of the experimental design. The proposed methods rely on specific dataset compositions, supervision levels, and modelling choices that necessarily delimit the generality of the conclusions. This section outlines the principal limitations related to scalability, genre leakage, evaluation design, and dataset bias, to clarify the conditions under which the reported results are expected to hold.

7.3.1 Scalability

A primary limitation of the approaches presented in this thesis concerns scalability across artists, datasets, and deployment settings. The biography corpora were constructed either through web-based collection or LLM generation, both of which are subject to long-tail effects. Popular artists tend to have richer and more detailed biographical information, whereas emerging or less-documented artists may be associated with sparse or low-quality descriptions. As a result, representation quality may vary systematically with artist popularity.

Scalability also arises at the modelling level. The empirical results suggest that the benefits of biography-based supervision are more pronounced in higher-data regimes, where sufficient labelled examples enable the model to align contextual artist-level information with task-specific targets. In lower-resource settings, the model may struggle to effectively integrate this additional semantic channel. Consequently, the observed improvements may not transfer uniformly across tasks or domains with limited annotated data.

Together, these factors highlight that, while biography-based supervision is effective under controlled conditions, its scalability depends on coverage and data availability.

7.3.2 Genre leakage

A potential limitation of biography-based supervision is the presence of explicit genre descriptors in artist biographies. Because genre labels frequently appear in biographical texts, there exists a risk that improvements in genre classification tasks may partially reflect direct label leakage rather than deeper semantic

integration. This concern is particularly relevant when textual representations are provided at inference time.

To assess this effect, we conducted ablation studies in which explicit genre terms were masked from the biographies. While performance decreased relative to the unmasked condition, improvements over audio-only baselines persisted. This indicates that explicit categorical cues contribute to the observed gains but do not fully account for them.

These findings suggest that biography-based representations operate at two levels. At one level, they may provide direct task-relevant information through explicit descriptors. At another level, they encode broader stylistic and contextual attributes (influences, instrumentation, cultural associations) that correlate with genre categories without explicitly naming them. Masking experiments reduce the most direct form of leakage, but they do not eliminate the possibility of indirect correlations between contextual language and genre labels.

Accordingly, while the results support the usefulness of biography-based augmentation, they should be interpreted with the understanding that genre classification tasks are particularly sensitive to textual shortcuts. Fully disentangling explicit label injection from contextual semantic enrichment remains a challenging problem.

7.3.3 Evaluation constraints

The findings of this thesis are conditioned by the specific evaluation framework adopted. To isolate the contribution of textual augmentation, experiments were conducted using fixed model architectures and controlled training protocols. While this design enables systematic comparison of representations, it also constrains conclusions to the selected backbone models and training regimes. The interaction between biography-based supervision and alternative architectures or large-scale pretrained encoders remains an open question.

Additionally, the pretraining experiments rely on a generative reformulation of downstream tasks. Although this unified framework facilitates the study of inductive bias transfer, it introduces dependencies on decoding strategies and text generation performance. As a result, comparisons with standard discriminative baselines should be interpreted in light of this modelling choice.

Finally, the evaluation focuses on supervised semantic tasks, namely genre classification, emotion recognition, and music tagging. While these benchmarks provide a meaningful framework for representation learning, they do not encompass other MIR settings such as retrieval, recommendation, or zero-shot transfer. Consequently, the reported improvements should be understood as specific to the evaluated tasks.

7.3.4 Dataset bias

Artist biographies are not uniformly distributed across the musical landscape. As discussed previously, the richness of biographies is closely tied to artists' visibility, introducing a popularity bias in which well-established artists receive

more extensive and detailed coverage. Temporal and regional factors further shape this imbalance, as publicly available documentation tends to privilege contemporary, Western, and anglophone artists. Consequently, the textual representations derived from these sources may disproportionately reflect dominant cultural narratives within popular music.

Importantly, similar distributional biases are present in the downstream evaluation benchmarks. Genre classification datasets often emphasise canonical Western genre taxonomies, and emotion recognition corpora frequently rely on culturally specific interpretations of affect. As a result, both the textual supervision and the evaluation tasks may share overlapping cultural assumptions.

The alignment between these distributions raises the possibility that some observed performance gains reflect compatibility within a shared cultural framing. In other words, improvements may be amplified when biography-derived contextual signals and benchmark annotations encode similar genre conventions or affective norms. While this does not invalidate the empirical findings, it suggests that generalisation to underrepresented genres, regions, or cultural contexts remains an open question.

7.4 Future work

Future work can be broadly divided into near-future methodological extensions and longer-term conceptual directions for knowledge-aware MIR.

7.4.1 Short-term directions

A first direction for further exploration concerns the robustness of the observed improvements across alternative audio encoding architectures. While the experiments in this thesis demonstrate consistent gains when combining artist biographies with VGGish and Whisper-based representations, it remains unclear how these effects interact with other model backbones or larger-scale pretrained audio encoders. Systematic evaluation across architectures with varying capacity and pretraining regimes would clarify whether biography-based supervision provides complementary benefits independent of encoder choice. It remains an open question whether the observed improvements would persist when employing state-of-the-art audio encoders such as MERT [29]. Still, we hypothesise that the subjective nature of the contextual information provided by biographies would continue to yield complementary signals even in the presence of more powerful acoustic representations. However, the relative magnitude of gains may decrease as the audio encoder captures more of the relevant semantic structure on its own.

A second line of investigation involves enforcing stricter separation between pretraining and downstream evaluation. In particular, future work could consider fully disjoint artist splits to assess whether biography-driven pretraining induces transferable contextual representations beyond artist-specific memorisation. Such experiments would provide a more rigorous test of whether the

observed gains reflect generalised inductive bias rather than overlap in artist identity.

Finally, further study is warranted in low-resource and long-tail settings. The empirical results suggest that biography-based augmentation is most effective in higher-data regimes, raising questions about its robustness under sparse supervision. Evaluating performance in limited-data scenarios and across underrepresented artists would help determine the extent to which biography-derived context can compensate for reduced acoustic or annotation signal.

7.4.2 Longer-term perspectives

More broadly, this thesis contributes to an emerging line of research concerned with incorporating structured contextual knowledge into MIR systems. Moving beyond purely signal-driven inference, future models may increasingly integrate cultural and stylistic information as complementary sources of supervision. Rather than relying exclusively on large-scale pretraining, such approaches could combine data-driven representation learning with explicit contextual grounding derived from artist-level knowledge.

A further long-term direction concerns interpretability and controllability in multimodal MIR systems. The distinction explored in this thesis between explicit biography injection and internalised pretraining highlights differences in transparency and modularity. Explicit textual conditioning provides a controllable interface for influencing model behaviour, whereas internalised knowledge operates implicitly within the representation space. Future research could explore how such controllable semantic axes might be leveraged to guide stylistic inference, mitigate unwanted biases, or adapt systems to new musical contexts.

Since the completion of this work, the field has moved rapidly toward large-scale music-language foundation models. Systems such as MusicLM, together with the MusicCaps dataset released alongside it [1], as well as contrastive audio-text models in the CLAP family [17], have demonstrated impressive capabilities in text-conditioned generation and cross-modal retrieval. However, many of the foundational questions raised in this thesis remain open in these larger-scale settings: whether textual supervision contributes through shallow lexical alignment or deeper semantic grounding, how models behave on long-tail artists with limited web presence, and whether reported gains generalise beyond the distributions seen during training. The controlled, representation-centric evaluation approach explored here offers a complementary perspective to scale-driven progress, and the core issues of semantic grounding and coverage are likely to become more pressing as music-language models increase their requirement of larger datasets.

Taken together, continued investigation of structured contextual supervision may contribute to more semantically grounded and culturally aware music representations, while preserving the benefits of data-driven learning.

7.5 Research impact

The primary impact of this thesis follows from its principal contribution, stated at the opening of this chapter: the empirical demonstration that artist-biography augmentation improves multimodal music representations, established under controlled conditions and defended against the most obvious alternative explanation through leakage-controlled ablation. The significance of this result is not the size of the improvement on any one benchmark, but what it indicates about where usable signal for these tasks resides. Genre, mood, and stylistic lineage are in substantial part culturally constructed categories, conventions about how music is described and grouped rather than properties that can be read off the waveform in the way tempo or spectral centroid can. That biographical text improves prediction of exactly these attributes, and continues to do so after explicit label terms are masked out, is evidence that high-level contextual knowledge complements acoustic modelling rather than duplicating it.

The practical implication is a direction for system design that runs alongside, rather than against, the prevailing one. The dominant response to difficulty in this area has been to scale the signal: more audio, larger encoders, more pretraining. This work suggests a complementary axis, building domain knowledge into the pipeline explicitly, whether injected at inference time or internalised through pretraining, and shows that it yields measurable returns at a scale accessible to academic research. It is offered as a complementary design consideration rather than a replacement for scale, and, as noted above, as a direction the evidence motivates rather than a settled result: which knowledge matters most, and how it is best introduced, remain open.

Beyond this central finding, the thesis contributes to Music Information Retrieval research by systematically examining the role of structured textual augmentation in music representation learning. Through a controlled and reproducible evaluation framework, it isolates the effects of artist-level contextual supervision and clarifies how such information interacts with acoustic representations. In doing so, the work provides methodological guidance for studying representation-level interventions in multimodal MIR systems.

A central practical contribution of this thesis is making available two corpora of artist biographies, constructed through web-based collection and LLM generation. Both are accessible for reuse, the web-scraped corpus as metadata together with a documented reconstruction procedure, since its source licences do not permit the biography text itself to be redistributed. By lowering the barrier to multimodal experimentation in academic settings, where access to proprietary large-scale datasets is often limited, this reproducible and high-coverage textual supervision supports further investigation into contextual augmentation strategies across tasks and architectures.

At a conceptual level, this thesis advances understanding of how contextual knowledge can influence learned representations of music. By distinguishing between explicit knowledge injection at inference time and inductive bias internalisation through pretraining, the work articulates two complementary mechanisms for integrating structured information into MIR systems. While the

reported gains are constrained by the experimental setup, the findings demonstrate measurable, consistent effects of contextual supervision, reinforcing the relevance of knowledge-aware approaches in multimodal learning.

Taken together, these contributions provide both practical resources and conceptual clarity for future research exploring semantically grounded music representations.

7.6 Conclusion

This thesis investigated whether structured artist-level textual knowledge can enhance music representation learning. To address this question, we developed a controlled evaluation framework designed to isolate the effects of representation choice and constructed two complementary corpora of artist biographies to enable systematic experimentation with contextual supervision.

Across semantic MIR tasks, the results demonstrate consistent improvements when incorporating biography-based representations, with the strongest effects observed in genre classification. The multimodal combined model reaches 89.3% accuracy on GTZAN (vs. 77.2% audio-only) and 72.9% on FMA-Small (vs. 52.7% audio-only). After controlling for explicit genre mentions through masking, these figures reduce to 78.3% and 71.1%, respectively, indicating that gains persist beyond surface-level lexical cues, particularly in larger-scale settings. For emotion recognition, multimodal fusion improves R^2 from 19.0 to 24.4 on Deezer, though EmoMusic remains inconclusive (R^2 46.6 vs. 51.4 audio-only), consistent with the challenges of low-resource affective prediction. For music tagging, improvements are more modest but consistent, with mAP increasing from 36.8 to 38.3 on MagnaTagATune and from 13.0 to 14.8 on Jamendo-MoodTheme. Together, these findings indicate that artist-level textual information provides complementary context for acoustic features, with the magnitude and reliability of gains varying across tasks and dataset conditions. Moreover, preliminary pretraining experiments suggest that such contextual knowledge can be internalised, influencing learned representations even when textual input is not available at inference time.

Taken together, these results support the conclusion that artist-level supervision constitutes a meaningful source of structured contextual information for music representation learning. The central contribution of this thesis is that demonstration itself: the controlled, ablation-defended evidence that biography augmentation improves multimodal MIR. The corpora, the evaluation framework, and the pretraining study are what made that demonstration possible, comparable, and extensible respectively, and they retain independent value; but the multimodal result is what the thesis is for. By additionally distinguishing between explicit knowledge injection and inductive bias internalisation, the work clarifies two complementary mechanisms through which external knowledge can shape multimodal MIR systems.

The appropriate weight to place on this conclusion should be stated as carefully as the conclusion itself. These findings are evidence towards the relevance

of contextual knowledge in music information retrieval, not proof of it, and artist biographies are one possible vehicle for embedding field knowledge into models rather than a uniquely suitable one. Bounded by the experimental scope described throughout this work, what the thesis establishes is that this vehicle carries real signal, that the signal is not an artefact of label leakage, and that it can be obtained reproducibly and at scale. On that basis, contextual grounding is a direction the evidence supports pursuing, and one whose value is likely to grow, not diminish, as music-language models come to depend on ever larger and less curated collections.

7.7 Reflections

At the outset of this thesis, my primary objective was to develop a multimodal approach capable of combining audio and text for music understanding tasks. Early in the project, however, I encountered a fundamental obstacle: the difficulty of aligning music and text in a way that is both semantically meaningful and practically reproducible. Unlike domains such as vision or speech, where large-scale aligned datasets are widely available, multimodal Music Information Retrieval is constrained by limited access to high-quality, redistributable resources. Much of my early work focused on attempting to work around this limitation, exploring different datasets and combinations of modalities in search of a viable path forward.

Over time, I came to recognise that this challenge was not merely a practical inconvenience, but a central aspect of the research problem itself. The issue was not only the lack of data, but the nature of the alignment between audio and textual descriptions of music. Existing textual resources in MIR tend to exhibit a trade-off between semantic richness and availability: lyrics are expressive but restricted, metadata is scalable but sparse, and large web-based corpora are often difficult to redistribute or control. This realisation shifted my perspective from attempting to assemble larger multimodal datasets toward explicitly studying how different forms of textual knowledge contribute to music representation learning under controlled conditions. The construction of artist biography corpora emerged not simply as a data engineering step, but as a way of making the problem itself more tractable and reproducible.

A second important lesson concerned experimental design. In the early stages of this work, I explored increasingly complex models and expanded the number of datasets and tasks under consideration, expecting that broader experimentation would yield clearer insights. In practice, the opposite was true: as the system grew in complexity, it became harder to understand which components were responsible for observed improvements. The decision to fix the downstream architecture and evaluate representations within a unified framework proved to be a turning point, enabling more interpretable comparisons and more defensible conclusions.

The empirical results also led to a more nuanced understanding of what multimodal improvements represent. While artist biographies improved perfor-

mance across most tasks, particularly for genre classification, further analysis revealed that these gains arise from a combination of mechanisms. Explicit lexical cues, such as genre terms appearing in biographies, contribute meaningfully to performance, but do not fully account for the observed improvements. Biographies also encode broader contextual information (stylistic influences, cultural associations, geographic context), which shapes model behaviour in less transparent ways. This led me to reconsider the interpretation of multimodal gains: improvements should not be taken as direct evidence of deeper musical understanding, but rather as the result of an interaction between surface-level signals and richer contextual knowledge. Disentangling these effects remains an important open challenge.

Perhaps the most significant conceptual development during this project was the distinction between using textual knowledge as an auxiliary modality at inference time and using it as a supervisory signal during training. Initially, I approached textual augmentation primarily as an input feature to be combined with audio representations. The exploration of generative pretraining reframed this perspective: textual knowledge can also act as a target that shapes a model's inductive biases, influencing how audio representations are learned even when text is absent at inference. This distinction became a central organising principle of the thesis and clarified two complementary pathways for incorporating contextual information into MIR systems.

More broadly, this work has led me to view multimodal MIR not simply as a problem of combining modalities, but as a problem of semantic grounding and evaluation. The central question is not only whether text improves performance, but what kind of knowledge it introduces, how that knowledge aligns with the target task, and under what conditions the resulting improvements can be trusted.

Finally, I believe the direction explored in this thesis has implications beyond the specific experiments presented here. As machine learning increasingly incorporates knowledge-rich supervision, the ability to integrate contextual information in a controlled and reproducible manner will become increasingly important. In MIR, where access to proprietary datasets remains a significant barrier, approaches that balance semantic richness with accessibility may play a key role in enabling broader participation and more transparent research.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Professor Roderick Murray-Smith and Professor Jeffrey Dalton, for their guidance, patience, and support throughout this PhD. Their complementary perspectives have shaped both this thesis and my development as a researcher.

This work was partially funded by MoodAgent, whose support made this research possible. I am grateful for the opportunity to work at the intersection of academic research and industry application.

I would also like to thank the members of the GRILL research group—Paul Owoicho, Sophie Fisher, Carlos Gemmell, and Iain Mackie—for the discussions, feedback, and camaraderie that made the day-to-day experience of research so much more enjoyable.

Finally, I owe a special thanks to Marta Zanetello for her unwavering support and patience through every phase of this PhD. This thesis would not have been possible without her.

Bibliography

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

[2] Pablo Alonso-Jimnez, Xavier Serra, and Dmitry Bogdanov. Music Representation Learning Based on Editorial Metadata from Discogs. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference*, pages 825–833, Bengaluru, India, December 2022. ISMIR. doi: 10.5281/zenodo.7316790. URL `https://doi.org/10.5281/zenodo.7316790`.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460, 2020.

[4] Thierry Bertin-Mahieux, Daniel Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 591–596, 2011.

[5] Dmitry Bogdanov, Alastair Porter, Philip Tovstogan, and Minz Won. MediaEval 2019: Emotion and theme recognition in music using Jamendo. In *MediaEval Benchmarking Initiative for Multimedia Evaluation*, 2019.

[6] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The MTG-Jamendo dataset for automatic music tagging. In *Proceedings of the Machine Learning for Music Discovery Workshop, ICML 2019*, 2019.

[7] Rodrigo Castellon, Chris Donahue, and Percy Liang. Codified audio language modeling learns useful representations for MIR. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, pages 82–89, 2021.

[8] Jeong Choi, Jongpil Lee, Jiyoung Park, and Juhan Nam. Zero-shot learning and knowledge transfer in music classification and tagging. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 67–74, Delft, The Netherlands, 2019.

[9] Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark B. Sandler. A tutorial on deep learning for Music Information Retrieval. *CoRR*, abs/1709.04396, 2017. URL `http://arxiv.org/abs/1709.04396`.

[10] Keunwoo Choi, Gyrgy Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2392–2396, 2017. doi: 10.1109/ICASSP.2017.7952585.

[11] Aurora Linh Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856, 2019. doi: 10.1109/ICASSP.2019.8682475.

[12] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 312–318, 2017.

[13] Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam. Music mood detection based on audio and lyrics with deep neural net. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*, pages 631–637, Paris, France, 2018.

[14] Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6964–6968, 2014. doi: 10.1109/ICASSP.2014.6854950.

[15] SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. LP-MusicCaps: LLM-based pseudo music captioning. In *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

[16] J. Downie, Kris West, Andreas Ehmann, and Emmanuel Vincent. The 2005 music information retrieval evaluation exchange (MIREX 2005): Preliminary overview. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*, pages 320–323, 2005.

[17] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[18] Josh Gardner, Simon Durand, Daniel Stoller, and Rachel M. Bittner. LLark: A multimodal instruction-following language model for music. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[19] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261.

[20] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[21] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017. doi: 10.1109/ICASSP.2017.7952132.

[22] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[23] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. MuLan: A joint embedding of music audio and natural language. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, pages 225–232, 2022.

[24] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.171.

[25] Edith Law, Kris West, Michael I. Mandel, Mert Bay, and J. Stephen Downie. Evaluation of Algorithms Using Games: The Case of Music Tagging. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pages 387–392, Kobe, Japan, October 2009. ISMIR. doi: 10.5281/zenodo.1417647. URL `https://doi.org/10.5281/zenodo.1417647`.

[26] Jongpil Lee and Juhan Nam. Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging. *IEEE Signal Processing Letters*, 24(8):1208–1212, 2017. doi: 10.1109/LSP.2017.2713830.

[27] Jongpil Lee, Jiyoung Park, Keunhyoung Kim, and Juhan Nam. Multi-level and multi-scale feature aggregation using pre-trained convolutional neural networks for music auto-tagging. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017*, pages 272–279, Suzhou, China, 2017.

[28] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Chenghua Lin, Xingran Chen, Anton Ragni, Hanzhi Yin, Zhijie Hu, Haoyu He, et al. MAP-Music2Vec: A simple and effective baseline for self-supervised music audio representation learning. In *Proceedings of the 24th International Society for Music Information Retrieval Conference, ISMIR 2023*, pages 642–649, Milan, Italy, 2023.

[29] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghua Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, et al. MERT: Acoustic music understanding model with large-scale self-supervised training. In *The Twelfth International Conference on Learning Representations*, 2024.

[30] Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. Music understanding LLaMA: Advancing text-to-music generation with question answering and captioning. *arXiv preprint arXiv:2308.11276*, 2023.

[31] Beth Logan. Mel Frequency Cepstral Coefficients for music modeling. In *Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR)*, 2000.

[32] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and Gyrgy Fazekas. Learning music audio representations via weak language supervision. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 456–460, 2022. doi: 10.1109/ICASSP43922.2022.9746996.

[33] Matthew McCallum, Korian Edeline, et al. Supervised and unsupervised learning of audio representations for music understanding. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022*, pages 535–542, Bengaluru, India, 2022.

[34] Sergio Oramas, Oriol Nieto, Mohamed Sordo, and Xavier Serra. A deep multimodal approach for cold-start music recommendation. In *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems*, DLRS 2017, page 3237, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450353533. doi: 10.1145/3125486.3125492. URL `https://doi.org/10.1145/3125486.3125492`.

[35] Sergio Oramas, Oriol Nieto, Mohamed Sordo, and Xavier Serra. Multi-label music genre classification from audio, text, and images using deep

features. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017*, pages 23–30, Suzhou, China, 2017.

[36] Sergio Oramas, Oriol Nieto, Francesco Barbieri, and Xavier Serra. Multi-modal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval*, 1(1):4–21, 2018. doi: 10.5334/tismir.10.

[37] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[38] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.200.

[39] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik M. Schmidt, Andreas F. Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 191–197, 2017.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[41] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023.

[42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

[43] Federico Rossetto, Jeffrey Dalton, and Roderick Murray-Smith. Generating multimodal augmentations with LLMs from song metadata for music information retrieval. In *Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications*, LGM3A '23, page 5159, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702839. doi: 10.1145/3607827.3616842. URL `https://doi.org/10.1145/3607827.3616842`.

[44] James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980. doi: 10.1037/h0077714.

[45] Mohammad Soleymani, Micheal N. Caro, Erik M. Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, CrowdMM '13, page 16, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450323963. doi: 10.1145/2506364.2506365. URL `https://doi.org/10.1145/2506364.2506365`.

[46] Janne Spijkervet and John Ashley Burgoyne. Contrastive learning of musical representations. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, pages 673–681, 2021.

[47] Bob L. Sturm. An analysis of the GTZAN music genre dataset. In *Proceedings of the Second International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*, MIRUM '12, page 712, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450315913. doi: 10.1145/2390848.2390851. URL `https://doi.org/10.1145/2390848.2390851`.

[48] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, et al. HEAR: Holistic evaluation of audio representations. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 125–145. PMLR, 2022.

[49] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002. doi: 10.1109/TSA.2002.800560.

[50] Julián Urbano, Markus Schedl, and Xavier Serra. Evaluation in music information retrieval. *Journal of Intelligent Information Systems*, 41(3): 345–369, 2013. doi: 10.1007/s10844-013-0249-4.

[51] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013. doi: 10.3758/s13428-012-0314-x.

[52] Felix Weninger, Florian Eyben, and Björn Schuller. On-line continuous-time music mood regression with deep recurrent neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5412–5416, 2014.

[53] Minz Won, Sanghyuk Chun, Oriol Nieto, and Xavier Serrc. Data-driven Harmonic Filters for audio representation learning. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 536–540, 2020. doi: 10.1109/ICASSP40776.2020.9053669.

[54] Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, et al. MARBLE: Music audio representation benchmark for universal evaluation. *Advances in Neural Information Processing Systems*, 36:39626–39647, 2023.

[55] Hang Zhao, Chen Zhang, Bo Zhu, Zeyu Ma, and Ke Zhang. S3T: Self-supervised pre-training with swin transformer for music classification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 606–610. IEEE, 2022.

[56] Jiahao Zhao, Ganghui Ru, Yi Yu, Yulun Wu, Dichucheng Li, and Wei Li. Multimodal music emotion recognition with hierarchical cross-modal attention network. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022. doi: 10.1109/ICME52920.2022.9859812.