Proceedings of Machine Learning Research 303:1–23, 2026 EAIM2026 at AAAI

Prevailing Research Areas for Music AI in the Era of Foundation Models

Megan Wei MEGANWEI@BROWN.EDU
Brown University

Mateusz Modrzejewski MATEUSZ.MODRZEJEWSKI@PW.EDU.PL
Warsaw University of Technology

Aswin Sivaraman ASIVARA@IU.EDU
Indiana University

Dorien Herremans DORIEN_HERREMANS@SUTD.EDU.SG
Singapore University of Technology and Design

Editors: D. Herremans, K. Bhandari, A. Roy, S. Colton, M. Barthet

Abstract

Parallel to rapid advancements in foundation model research, the past few years have witnessed a surge in music AI applications. As AI-generated and AI-augmented music become increasingly mainstream, many researchers in the music AI community may wonder: what research frontiers remain unexplored?

This paper outlines several key areas within music AI research that present significant opportunities for further investigation. We begin by examining foundational representation models and highlight emerging efforts toward explainability and interpretability. We then discuss the evolution toward multimodal systems, provide an overview of the current landscape of music datasets and their limitations, and address the growing importance of model efficiency in both training and deployment.

Next, we explore applied directions, focusing first on generative models. We review recent systems, their computational constraints, and persistent challenges related to evaluation and controllability. We then examine extensions of these generative approaches to multimodal settings and their integration into artists' workflows, including applications in music editing, captioning, production, transcription, source separation, performance, discovery, and education.

Finally, we explore copyright implications of generative music and propose strategies to safeguard artist rights. While not exhaustive, this survey aims to illuminate promising research directions enabled by recent developments in music foundation models.

1. Introduction

Music information retrieval (MIR) is a mature interdisciplinary research area that has seen significant progress across many fronts, from model architectures and advanced music tagging to broadening coverage of global musical cultures. However, the emergence of foundation models has transformed the landscape, capturing the attention of researchers, artists, and the public alike. Major news outlets are now covering AI-powered tools for music composition as well as AI-generated music (e.g., the "audio deepfakes" of high-profile artists) (Chow, 2023). Moreover, music AI startups such as Suno, Udio, and Producer have rapidly emerged, offering prompt-based music synthesis, acoustic style transfer, and symbolic music control. These tools have accelerated creative workflows, enabling artists to iteratively

© 2026 M. Wei, M. Modrzejewski, A. Sivaraman & D. Herremans.

explore new sounds and melodies as part of their ideation process. Rapid improvements in foundation models have achieved capabilities previously considered unattainable. This immense progress might suggest that most challenges in MIR and music AI have been solved. This is not quite true, in addition, these same advances introduce novel paradigms that reveal new research directions and unresolved questions, fundamentally reshaping what problems we should address.

In this work, we explore emerging research directions in MIR, emphasizing the need for rigorous validation and integrity. While space constraints prevent comprehensive coverage, we highlight critical open problems and opportunities across these three categories: fundamental, applied, and responsible music AI. Our goal is to help the community rethink what is possible in this era and inspire potential answers to the age-old researcher's dilemma: ***what should I work on now?***

2. Fundamental Music AI

2.1. Music encoders

In computer vision, ResNets (He et al., 2016) and vision transformer models (Dosovitskiy et al., 2020) have learned rich visual representations from extensive image datasets. In natural language processing, BERT (Devlin et al., 2018) demonstrated large-scale language understanding, while the T5 transformer (Raffel et al., 2020) introduced a unified text-to-text interface, paving the way for modern large language models (LLMs).

However, research on music foundation models has only recently gained momentum (Ma et al., 2024). A music foundation model would need to understand music in multiple dimensions – melody, harmony, rhythm, timbre, and diverse stylistic and cultural traits – yielding universal representations for many downstream tasks. The Holistic Evaluation of Audio Representations (HEAR) challenge at NeurIPS 2021 (Turian et al., 2022) was among the first large-scale initiatives towards robust audio representations, providing a benchmark of 19 diverse tasks.

Generatively pre-trained latent representations from models like Jukebox (Dhariwal et al., 2020) and MERT (Li et al., 2023) have proven useful for a variety of tasks, as demonstrated by evaluations on the MARBLE Benchmark (Yuan et al., 2023). Neural audio codecs such as EnCodec (Défossez et al., 2022), Descript (Kumar et al., 2023) and SoundStream (Zeghidour et al., 2021) employ discrete neural representation learning (Van Den Oord et al., 2017) to encode audio into discrete tokens, useful for generative tasks, like MusicGen (Copet et al., 2024) and MusicLM (Agostinelli et al., 2023). Finally, Won et al. (2024) and Gardner et al. (2023) augment foundation models with music understanding.

Developing a unified music generation and understanding pipeline as well as building more efficient and robust architectures remains an open and activate research question.

2.2. Explainability

Explainable artificial intelligence (XAI), essential for developing trustworthy and interpretable systems, has seen limited adoption in MIR. Early work by Choi et al. (2016); Won et al. (2019) visualized and auralized spectrogram features to explain music tagging models, analogous to saliency maps in image processing. Foscarin et al. (2022) introduced a

concept-based approach for explaining composer classification, that prioritized interpretability for musicologists rather than solely AI experts. Mishra et al. (2017) adapted the LIME framework (Ribeiro et al., 2016) to music content analysis and later to singing voice detection (Mishra et al., 2020), while Haunschmid et al. (2020) proposed audioLIME, which replaces image segmentation with source separation. Despite these efforts, adaptation of broader explainable artificial intelligence (XAI) tools such as Shapley additive explanations (SHAP) (Lundberg and Lee, 2017), layerwise relevance propagation (LRP) (Bach et al., 2015), and concept relevance propagation (CRP) (Achtibat et al., 2023) have seen sparse adoption, particularly beyond music tagging.

A newer direction, driven by generative models focuses on explanation through data attribution, identifying which musical content shapes system outputs. Barnett et al. (2024b) study embedding similarities between prompts and outputs in VampNet (Garcia et al., 2023), while Batlle-Roca et al. (2024) propose metrics and tools to detect data replication.

2.3. Interpretability

Several studies in language modeling have examined whether model representations encode meaningful features useful for downstream tasks (Meng et al., 2022; Tenney et al., 2019), spanning domains from color (Abdou et al., 2021) and world knowledge (Li et al., 2022; Yun et al., 2023) to auditory perception (Ngo and Kim, 2024). Interpreting these internal representations offers an alternative to traditional MIR tagging approaches, helping to overcome limited dataset annotations and enabling potential inference-time control through representation editing. Recent work training probe classifiers on embeddings of music generative models shows that they capture both high-level musical concepts—such as genre, emotion, and instrument type (Castellon et al., 2021; Koo et al., 2024)—and lower-level music-theoretic features (Wei et al., 2024). Other studies explore how architecture and self-supervised learning influence such representations in music understanding tasks (Won et al., 2023; Li et al., 2023). Meanwhile, several works (Zhao et al., 2025; Singh et al., 2025) have leveraged internal representations of musical concepts for activation steering (Turner et al., 2024). Future research may further explore disentanglement in polysemantic features (Elhage et al., 2022), where individual neurons in a music generation models may be entangled with multiple musical features.

2.4. Multimodality

There is a clear trend toward multimodal generative AI. While Section 3.1 discusses the rise of text-to-music systems, video-to-music generation remains relatively unexplored (Kang et al., 2024; Su et al., 2024). Similarly, real-time, adaptive music generation presents promising opportunities for conditioned generation, though it faces challenges such as latency constraints and limited training datasets.

Joint embedding models, which represent multiple modalities in a shared semantic space, facilitate cross-modal retrieval, alignment scoring, and generation. These models are typically trained via contrastive learning, as in the vision-language model CLIP (Radford et al., 2021), and analogous audio-text models like CLAP (Contrastive Language–Audio Pretraining) (Elizalde et al., 2023; Wu et al., 2023c), MuLan (Huang et al., 2022), and text-MIDI models such as CLaMP (Wu et al., 2023a).

Despite these advances, multimodal representation learning for symbolic music remains underexplored. Recent work also shows that multimodal audio models can over-rely on language, potentially compromising audio fidelity (Weck et al., 2024). Future research may benefit from integrating symbolic, audio, video, and textual modalities, enabling cross-modal transformations and leveraging larger datasets to build more robust foundational representations.

2.5. Efficiency

Efficiency has become a central concern as music AI models grow in size and complexity. Smaller, faster models are not only more accessible to researchers with limited computational resources, but are also essential for latency-sensitive applications such as live music performance and interactive systems. Recent progress has been driven by techniques such as mixed-precision training, model quantization, and specialized hardware acceleration.

Lightweight models combining signal processing and deep learning have emerged as an active research area. PESTO (Riou et al., 2023) achieves state-of-the-art fundamental frequency estimation with only 29k parameters by leveraging the transpositional invariance of the Constant-Q Transform (CQT), while Basic Pitch (Bittner et al., 2022) performs accurate pitch tracking with around 16k parameters using a harmonic CQT representation. RAVE (Caillon and Esling, 2021) enables real-time audio synthesis and style transfer, operating up to 80× faster than real time on a standard CPU through multi-band decomposition and Pseudo Quadrature Mirror Filters (PQMF) (Nguyen, 1994). Even in the text-to-song domain, relatively tiny models such as JAM (Liu et al., 2025) demonstrate that optimized architectures and high-quality data can efficiently achieve strong performance.

Many of the most efficient models achieve their performance by integrating of Digital Signal Processing (DSP) methods—such as Fourier variants and audio decomposition—with modern deep learning paradigms. Continued research at this intersection is likely to remain key for advancing efficient, real-time music AI systems. The next section discusses several applied areas that have seen rapid development with the rise of AI technologies.

3. Selected music AI applications

We discuss a handful of selected applications that benefit from the recent developments in AI. We will briefly discuss a few of the latest systems as well as the challenges remaining and opportunities offered by these foundation models.

3.1. Generative Models

While comprehensive surveys of generative music models exist (Herremans et al., 2017; Ji et al., 2023; Le et al., 2024; Chen et al., 2025), we briefly highlight recent systems and remaining challenges.

Text-conditioned audio-based generation has advanced rapidly (Melechovsky et al., 2024a; Agostinelli et al., 2023; Copet et al., 2024), yet many still generate only short or instrumental-only fragments, though cascading approaches enable longer outputs (Schneider et al., 2023). Open-source models are limited, with notable examples including MusicGen (Copet et al., 2024), Stable Audio Open (Evans et al., 2024), Magenta Real-Time (Caillon et al., 2025),

and Mustango (Melechovsky et al., 2024a); even fewer are trained solely on open data. Many commercial models, such as those from Suno and Udio, are unavailable for research and may rely on copyrighted datasets.

Real-time music generation (Caillon et al., 2025) in particular with human interaction (Blanchard et al., 2025; Scarlatos et al., 2025) represents an exciting avenue for artist-in-the-loop musical performance. However, quality real-time generation remains constrained by hardware requirements. High-quality systems typically require GPUs—for instance, Mustango takes 40 seconds to generate 10s of audio on an NVIDIA A100. Research into GPU-optimized architectures and efficiency improvements, such as the flow-based model JAM (Liu et al., 2025), is crucial, especially with the rise of inference alignment techniques (Herremans and Roy, 2025; Roy et al., 2025b). Furthermore, developing models that can run on-device in real-time, responsive to human interactivity, would greatly democratize music co-creation.

For practical adoption, generative tools must integrate with musicians' familiar digital audio workstations (DAWs), typically via virtual studio technology (VST) plugins (Marrington et al., 2017). However, this integration remains challenging because VSTs are coded in C++ while most machine learning models use Python. Limited attempts have been made to bridge the gap between Python and VSTs (Braun, 2021). Such interfaces would enable these systems to become true tools for co-creation.

Singing voice synthesis presents another significant challenge, primarily due to its demand for large amounts of copyright-free training data. Recent models like JAM (Liu et al., 2025), DiffRhythm+ (Chen et al., 2025), SongBloom (Yang et al., 2025), YuE (Yuan et al., 2025), and Ace Step (Gong et al., 2025) demonstrate progress, but still require massive datasets (e.g., DiffRhythm+ trained on over one million songs), which are rarely openly available.

Finally, MIDI remains a popular format among professional music producers in DAW software. Current work for text-to-MIDI generation include MuseCoco (Lu et al., 2023) and Text2midi (Bhandari et al., 2025), while (Donahue et al., 2024) leverages the HookTheory platform for human-AI co-creation in symbolic music. Many more symbolic music datasets have emerged, such as MidiCaps (Melechovsky et al., 2024b) and Aria (Bradshaw and Colton, 2025) (transcribed audio recordings). However, high-quality data remains scarce, limiting the capabilities of generative MIDI models as discussed in Section 4.1. In the following subsections, we dive deeper into the various aspects of music generation systems.

3.1.1. Evaluation Metrics

Evaluating music and audio systems remains challenging due to the subjective nature of human perception and the diversity of listener preferences. While numerous metrics exist for both audio (e.g., KL divergence, Fréchet Audio Distance (Kilgour et al., 2018)) and symbolic systems (Dong et al., 2020), the lack of standardization across studies makes benchmarking and comparison difficult. Mauve Audio Divergence (Huang et al., 2025) seeks to improve previous systems like FAD by better aligning evaluation metrics with human preferences through pretrained foundation model representations.

To assess semantic alignment between music and text prompts, metrics like CLAP-score (Elizalde et al., 2023) are commonly used, though CLAP was not specifically trained for music-theory and finer-grained concepts. Alternative approaches have relied on feature

predictors to compare generated music with input prompts (Melechovsky et al., 2024a). In the symbolic domain, metrics such as musical interval statistics (Makris et al., 2021) and structure-based measures like COSIATEC compression ratio (Meredith, 2013) are employed (Herremans and Chew, 2017). For a comprehensive overview of objective evaluation metrics, readers are referred to Ji et al. (2020); Kader and Karmaker (2025).

Despite these efforts, objective metrics alone are often insufficient (Vinay and Lerch, 2022), and a growing consensus suggests that deeper musical insight is needed (Modrzejewski and Rokita, 2023). As a result, subjective evaluation through listener surveys remains essential, though these require careful design to ensure meaningful results. Standardized listening tests are needed, and practices like cherry-picking examples should be avoided. The recent Music Arena platform (Kim et al., 2025) allows for users to generate and rate music across different models. Agres et al. (2016) offer a comprehensive survey of subjective evaluation techniques for creative music systems.

Recent work has shifted from low-level signal or embedding metrics toward measures that better capture perceived musical quality. SongEval (Yao et al., 2025) introduces a dataset of over 2,000 full-length songs annotated by professional musicians across dimensions such as coherence, phrasing, and overall musicality. Similarly, Meta's Audiobox-Aesthetics model (Tjandra et al., 2025) predicts perceived audio quality across speech, music, and environmental sounds, assessing attributes like production quality, complexity, and enjoyment. Together, these models move evaluation toward human-centered aesthetic judgment rather than purely objective or distributional measures. Their quantitative nature also enables integration into Direct Preference Optimization (DPO) (Dong et al., 2024) or inference-time output selection frameworks (Roy et al., 2025b), where aesthetic scores can guide ranking or fine-tuning toward more musically appealing results.

3.1.2. Controllability

Controllable generation aims to guide AI models toward outputs with user-specified attributes and has seen progress in music AI. Traditional MIDI-based models allow conditioning on musical features or emotions (Makris et al., 2021), while recent text-based models primarily provide high-level control (e.g., genre, mood). Only a few models, such as Mustango, enable incorporation of music-theoretic constraints, allowing generation with specific chord sequences, tempos, or keys.

Text control is typically global, but fine-grained, time-varying control over low-level aspects like dynamics or tempo requires more intuitive interfaces. Music ControlNet (Wu et al., 2023b) addresses this by combining global controls (mood, genre) with time-varying controls (melody, rhythm, dynamics), while other works adopt symbolic, compositional, or rule-based methods (Huang et al., 2024; Thickstun et al., 2023).

Emotion is another key dimension of control. MIDI-based models have explored emotion conditioning (Herremans and Chew, 2017; Makris et al., 2021). While LLM-based systems often account for emotion implicitly, explicit emotion control remains underexplored, yet it holds potential for applications in mental health and personalized music generation (Agres et al., 2021).

Iterative music editing remains a promising direction for interactive, controllable music co-creation. Models such as MusicMagus (Zhang et al., 2024b) and Instruct-MusicGen (Zhang et al., 2024a) are equipped with large multimodal models that allow the user to

change the instrument, genre, mood, or add an additional stem on top of the existing accompaniment. To train such models to generate and edit musical stems, such as MusicGen-Stem (Rouard et al., 2025), researchers often build upon datasets for source separation, which can be augmented for stem generation.

Looking forward, substantial room for improvement remains in enabling musicians and composers to truly co-create with AI through more descriptive, abstract, temporally aware, and theory-informed instructions, e.g. 'music for a gym workout, starting with an ascending melody, introducing a key change up two semitones, and ending on a C-note.'.

3.2. Music production and mixing

Recent advances in AI-based music production have introduced new approaches to automated music mastering (AMM) (Steinmetz et al., 2021; Martínez-Ramírez et al., 2022). However, progress is limited by the lack of publicly available training data—namely, paired dry (anechoic) recordings and their mastered counterparts. Expert annotations from professional audio engineers could further improve supervised learning methods. Moreover, several studies note the shared objectives between music source separation and AMM, suggesting that joint optimization could yield more tractable and interpretable mixing and separation strategies (Yang et al., 2022).

Recently, text-controlled music production models have emerged. Notably, Chu et al. (2025)'s text2fx introduced text-guided audio effects processing, while SonicMaster (Melechovsky et al., 2025), a flow-based model, represents the first all-in-one text-driven system for music restoration and mastering. To overcome the lack of training data, SonicMaster artificially degrades production-quality recordings using 16 types of distortions, enabling it to respond to natural-language commands such as "bring out the bass" or "reduce the reverb."

Beyond mixing and mastering, rendering human-like expressiveness from MIDI remains a key challenge. This spans expressive performance generation from quantized MIDI (Cancino-Chacón et al., 2018) and the design of neural sound fonts, from models emulating analog gear such as tube amplifiers (Damskägg et al., 2019) to those directly synthesizing instrument sounds from symbolic input (Castellon et al., 2020). Although research in this area has slowed, there remains a clear need for end-to-end systems that transform MIDI into expressive, fully mastered audio by unifying synthesis and mastering.

3.3. Music Captioning

Music captioning–the task of generating natural-language descriptions of musical audio–is also emerging as a challenge in music AI. Early work, such as MusCaps (Manco et al., 2021) introduced encoder–decoder architectures that produce a descriptive sentence (e.g., instrumentation, mood, genre) rather than classical tags. Recently, MusiLingo (Deng et al., 2023) bridges pretrained music-audio encoders (e.g., MERT) with large language models (LLMs) via a projection layer, enabling music caption generation and query-response interactions ("what is the mood?", "which instruments?", etc.). SonicVerse (Chopra et al., 2025) advances the field by adopting a multi-task architecture: besides caption generation, it jointly predicts auxiliary musical feature labels (key, vocal presence, instrumentation, genre) and injects these feature token projections into a language model to produce richer captions. Meanwhile, large multimodal audio-language models such as Qwen3-Omni (Xu et al., 2025)

suggest the potential for generalized audio captioning systems that can handle music, speech and ambient sound in a unified way, though their specialization for music captioning remains underexplored.

Despite this progress, considerable challenges remain. Datasets are still small relative to those in vision or language domains, and standard captioning metrics (BLEU, ROUGE, METEOR) correlate poorly with human judgments in the music domain (Lee and Lee, 2024). Furthermore, musical audio presents overlapping layers (multiple instruments, effects, production artifacts, and expressive timing) and a time element that complicate capturing high-level semantics in text. Overcoming these bottlenecks and scaling captioning systems across large song catalogs could enable richly annotated natural-language descriptions that improve both retrieval and the training of generation systems.

3.4. Music transcription

Music transcription—the automatic conversion of audio recordings into symbolic representations such as MIDI—is often described as a "holy grail" of the MIR field. While considerable progress has been made, particularly for monophonic recordings or isolated instruments (e.g., note-level transcription of solo piano), accurate transcription of polyphonic, multi-instrument performances remains a major challenge. For instance, Wang et al. (2018) proposed a note-based music language model combined with an acoustic front-end for polyphonic piano transcription, while Simon et al. (2022) demonstrated that scaling training data via mixtures of monophonic recordings improved multi-instrument polyphonic transcription. Overcoming this barrier would enable transcription of millions of existing audio songs into accurate, richly annotated polyphonic MIDI (or equivalent symbolic) files. Such a dataset would be a transformative resource, enabling the training of large-scale generative MIDI models such as Text2midi (Bhandari et al., 2025).

3.5. Source separation

In recent years, neural networks have achieved state-of-the-art performance in music source separation (Stoller et al., 2018; Défossez et al., 2019). Models such as Spleeter (Hennequin et al., 2020) have been open-sourced, while proprietary tools like LALAL.AI have seen widespread commercial adoption. Recent research has explored novel model architectures such as band-split RNN (Luo and Yu, 2023), transformers (Rouard et al., 2023), as well as new training paradigms (e.g., augmenting training data (Pons et al., 2024) or modifying the loss function (Sawata et al., 2023)). Despite these advances, separating certain instruments, such as distinct singing voices or synthesizers, remains challenging. Accurate source separation would not only empower music producers but also enable the generation of large-scale stem datasets for training music generation models.

3.6. Music Discovery

The integration of recommender systems, deep learning, and music information retrieval has transformed how users discover music, enabling highly personalized suggestions based on listening habits and preferences. Yet this personalization introduces challenges, such as the echo chamber effect, where algorithms reinforce existing tastes and limit exposure to new genres, artists, or cultures. Long-term user engagement remains difficult to capture

(Wang et al., 2022), while recent surveys emphasize the growing importance of fairness and transparency in recommender design (Wang et al., 2023). Zangerle and Bauer (2022) highlight the need for long-tail recommendations that promote older or lesser-known tracks, and Knees et al. (2020) analyze discovery interface principles that balance metadata quality with contextual and artist-centric cues.

Recent advances extend beyond conventional collaborative filtering toward multimodal and conversational recommendation. Models such as CLAP (Elizalde et al., 2023) and MuLan (Huang et al., 2022) align text and audio representations, allowing users to search and discover music through natural-language descriptions rather than fixed tags. Emerging LLM-driven recommenders (e.g., ChatMusicRec (Kim et al., 2024), LLM-Rec (Liu et al., 2023)) act as dialogue agents that elicit user intent and justify their suggestions, providing more transparent and exploratory discovery experiences.

3.7. Music Performance

Using neural networks as live musical instruments introduces a new paradigm in musical creativity, where artificial intelligence collaborates with human musicians to generate novel sounds and compositions. For instance, RAVE (Caillon and Esling, 2021) has been used to perform drum rhythms through integration with Max[1] and motion sensor controls, allowing to create rhythms through expressive gestural movements in the air. Recently, researchers have developed an AI symbolic music jam bot, based off of Anticipatory Music Transformer (Thickstun et al., 2023), to perform a live concert in real-time with GRAMMY-winning artist Jordan Rudess (Blanchard et al., 2025). The sampling and looping capabilities of VampNet (Garcia et al., 2023) have been used for creating live experimental soundscapes with a live instrumentalist through dynamic interplay between the musician with the model.

AI models can also adapt in real-time to human performer inputs, enabling otherwise physically impossible instruments. The HITar exemplifies this approach: an augmented acoustic guitar played with hybrid-percussive technique, morphed and extended in real-time by AI-generated tabla sounds (Martelloni et al., 2023). We consider the usage of AI to create entirely new, previously undiscovered means of musical expression an especially exciting and worthwhile endeavor.

3.8. Music Education

Although applications of AI in music education remain under-explored, most approaches aim either to assist students directly in musical practice or to enhance the accessibility and therapeutic aspects of learning music. These include systems that provide feedback, generate personalized learning materials, or offer new modes of musical interaction. Gover and Zewi (2022) propose difficulty-level conversion, generating easier or harder versions of piano arrangements, while pitch and beat tracking models can support vocal or instrumental practice. Morsi et al. (2023) detect conspicuous mistakes in piano performances without requiring a score for comparison, and Balliauw et al. (2017) automate the generation of piano fingerings for students.

More recently, large language models and multimodal AI systems have been explored as intelligent tutors and performance coaches. Pond and Fujinaga (2025) show that LLMs

1. https://cycling74.com/products/max

can reason about music theory through in-context and chain-of-thought prompting, while Zhang et al. (2025) propose LLaQo, a multimodal coach providing expressive feedback on performance. Together, these studies suggest that adaptive, conversational, and multimodal AI systems could soon play a central role in personalized and creative music learning.

4. Responsible Music AI

4.1. Datasets

The diversity of MIR tasks creates significant opportunities for developing new datasets to advance the field. Historically, audio datasets have varied in format: some provide small collections of raw audio, while others distribute only extracted features or track IDs (Won et al., 2021). MIDI datasets represent another important modality, enabling research on note sequences, timing, and musical structure, but often require different processing approaches than audio. While it is impossible to discuss all MIR datasets comprehensively, we outline several key challenges below.

Musical stems are a particularly sought-after data source, as music producers naturally work with stems in their DAW, and such datasets can facilitate the training of stem-generation and source separation models. MIDI datasets can be synthesized per track, enabling stem-separated audio. However, rendering MIDI introduces significant artifacts and limited expressivity (Manilow et al., 2019). Real instrument datasets like MUSDB18 (Rafii et al., 2017), though popular, are relatively small (approximately 10 hours) and now serve more as benchmarks than as comprehensive training resources.

To address coverage gaps in instruments and genres, researchers increasingly turn to larger open-domain datasets such as FMA, which offers over 106,000 tracks or roughly 343 days of Creative Commons-licensed audio. However, these often lack individual instrument stems, requiring self-supervised or unsupervised methods for MSS.

A major limitation persists in legally obtainable training data for text-to-music generation. Captioned music datasets are scarce, with only a few copyright-cleared options available—MusicCaps (Agostinelli et al., 2023), MusicCaps-LM (Doh et al., 2023), MusicBench (Melechovsky et al., 2024a), and JamendoMaxCaps (Roy et al., 2025a). For MIDI, the recently released MidiCaps (Melechovsky et al., 2024b) is the only dataset offering full-length files with detailed text descriptions. This scarcity has limited progress in text-to-MIDI systems, though recent models like MuseCoco (Lu et al., 2023) attempt to overcome these gaps by extracting features from input captions. In all these datasets, it is important to also consider cultural (particularly Western) bias as well as the annotation quality (e.g. LLM-annotated captions versus human-annotated).

In summary, dataset limitations remain a significant bottleneck. There is a substantial need for copyright-cleared stem and MIDI datasets, annotated with text captions including music-theory descriptions (to enable reasoning) and general descriptions such as instrumentation, emotion, and mood.

4.2. Respecting artist rights

Over the years, disruptive technologies—from cable television to MP3 players and streaming services—have repeatedly clashed with copyright holders, often resulting in legal battles

that ultimately favored technological progress. Today, generative AI models are advancing at a pace that far exceeds the evolution of laws and policies (Samuelson, 2023). Recent litigation has focused on large language models reproducing copyrighted content, such as song lyrics, from both licensed and web-scraped training data. For instance, Anthropic was sued by major music publishers for alleged infringement, while organizations like ASCAP are advocating for legislation that prioritizes human creators, consent, compensation, credit, transparency, and global consistency (Poritz, 2023).

Copyright remains a central concern, as using licensed music tracks for model training–especially for commercial purposes—poses significant legal and ethical challenges. While copyright protections support artists and labels, they can also hinder the development of new music technologies. There is a growing need for equitable, easy-to-apply attribution mechanisms for music data and metadata (Herremans, 2025). Research into frameworks for attribution, fairness, and openness evaluation is increasingly important to balance the interests of creators and technologists. Potential solutions include robust melody and timbre similarity detection algorithms, which could facilitate legal data usage and profit-sharing without stifling research Herremans (2025); Lu et al. (2025).

In addition, deep learning-based watermarking (Roman et al., 2024; Chen et al., 2023), originally designed for speech, is now being adapted for music. Data poisoning methods such as Glaze (Shan et al., 2023) and Nightshade (Shan et al., 2024) have also shown promising results in the image domain, by applying subtle modifications to images that render the data ineffective for training specific generative models. Early research has begun exploring similar techniques for music (Barnett et al., 2024a). The development of new technologies for protecting artist rights is emerging as an important research topic.

5. Conclusion

So what should I work on now? Although the history of AI and music extends as far back as 1950s, algorithms developed in just the last few years have achieved unparalleled capabilities, amplifying the now mainstream dialogue about AI. The technologies and software discussed in this survey have already begun to impact music educators, producers, performers, record labels, and artists at every level.

Our discussion reveals that this rapid progress has introduced new challenges and opportunities unique to the field of MIR. From ethical datasets and copyright attribution, to multimodal controllable systems, real-time generative modeling, and novel methods for music production, remixing, and editing, there has never been a better time to become a music AI researcher.

6. Acknowledgments

This work has received support from SUTD's Kickstart Initiative under grant number SKI 2021_04_06 and MOE T2. We acknowledge the use of ChatGPT for grammar improvements.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color. In Arianna Bisazza and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online, November 2021. Association for Computational Linguistics.

R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, and S. Lapuschkin. From Attribution Maps to Human-Understandable Explanations through Concept Relevance Propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023.

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.

K. R. Agres, J. Forth, and G. Wiggins. Evaluation of Musical Creativity and Musical Metacreation Systems. *Computers in Entertainment (CIE)*, 14(3):1–33, 2016.

Kat R Agres, Rebecca S Schaefer, Anja Volk, Susan van Hooren, Andre Holzapfel, Simone Dalla Bella, Meinard Müller, Martina De Witte, Dorien Herremans, Rafael Ramirez Melendez, et al. Music, computing, and health: a roadmap for the current and future roles of music technology for health care and well-being. *Music & Science*, 4:2059204321997709, 2021.

S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On Pixel-Wise Explanations for Non-linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, 2015.

Matteo Balliauw, Dorien Herremans, Daniel Palhazi Cuervo, and Kenneth Sörensen. A variable neighborhood search algorithm to generate piano fingerings for polyphonic sheet music. *International Transactions in Operational Research*, 24(3):509–535, 2017.

Julia Barnett, William Agnew, Robin Netzorg, Patrick O'Reilly, Ezra Awumey, Chris Donahue, and Sauvik Das. Audio data defenses: Protecting music and speech data from targeted attacks. In *Extended Abstracts for the Late-Breaking Demo Session of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024a.

Julia Barnett, Hugo Flores Garcia, and Bryan Pardo. Exploring musical roots: Applying audio embeddings to empower influence attribution for a generative music model. *arXiv preprint arXiv:2401.14542*, 2024b.

Roser Batlle-Roca, Wei-Hisang Liao, Xavier Serra, Yuki Mitsufuji, and Emilia Gómez. Towards assessing data replication in music generation with music similarity metrics on raw audio. *arXiv preprint arXiv:2407.14364*, 2024.

Keshav Bhandari, Abhinaba Roy, Kyra Wang, Geeta Puri, Simon Colton, and Dorien Herremans. Text2midi: Generating symbolic music from captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23478–23486, 2025.

Rachel M Bittner, Juan José Bosch, David Rubinstein, Gabriel Meseguer-Brocal, and Sebastian Ewert. A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 781–785. IEEE, 2022.

Lancelot Blanchard, Perry Naseck, Stephen Brade, Kimaya Lecamwasam, Jordan Rudess, Cheng-Zhi Anna Huang, and Joseph Paradiso. The jam_bot, a real-time system for collaborative free improvisation with music language models. In *ISMIR*, 2025.

Louis Bradshaw and Simon Colton. Aria-midi: A dataset of piano midi files for symbolic music modeling. *arXiv preprint arXiv:2504.15071*, 2025.

David Braun. Dawdreamer: Bridging the gap between digital audio workstations and python interfaces. *arXiv preprint arXiv:2111.09931*, 2021.

Antoine Caillon and Philippe Esling. Rave: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011*, 2021.

Antoine Caillon, Brian McWilliams, Cassie Tarakajian, Ian Simon, Ilaria Manco, Jesse Engel, Noah Constant, Pen Li, Timo I. Denk, Alberto Lalama, Andrea Agostinelli, Anna Huang, Ethan Manilow, George Brower, Hakan Erdogan, Heidi Lei, Itai Rolnick, Ivan Grishchenko, Manu Orsini, Matej Kastelic, Mauricio Zuluaga, Mauro Verzetti, Michael Dooley, Ondrej Skopek, Rafael Ferrer, Zalán Borsos, Äaron van den Oord, Douglas Eck, Eli Collins, Jason Baldridge, Tom Hume, Chris Donahue, Kehang Han, and Adam Roberts. Live music models. *arXiv:2508.04651*, 2025.

Carlos E Cancino-Chacón, Maarten Grachten, Werner Goebl, and Gerhard Widmer. Computational models of expressive music performance: A comprehensive and critical review. *Frontiers in Digital Humanities*, 5:25, 2018.

Rodrigo Castellon, Chris Donahue, and Percy Liang. Towards realistic midi instrument synthesizers. In *NeurIPS Workshop on Machine Learning for Creativity and Design*, 2020.

Rodrigo Castellon, Chris Donahue, and Percy Liang. Codified audio language modeling learns useful representations for music information retrieval. July 2021.

Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. Wavmark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*, 2023.

Huakang Chen, Yuepeng Jiang, Guobin Ma, Chunbo Hao, Shuai Wang, Jixun Yao, Ziqian Ning, Meng Meng, Jian Luan, and Lei Xie. Diffrhythm+: Controllable and flexible full-length song generation with preference optimization. *arXiv preprint arXiv:2507.12890*, 2025.

K. Choi, G. Fazekas, and M. Sandler. Explaining Deep Convolutional Neural Networks on Music Classification. *arXiv preprint arXiv:1607.02444*, 2016.

Anuradha Chopra, Abhinaba Roy, and Dorien Herremans. Sonicverse: Multi-task learning for music feature-informed captioning. *arXiv preprint arXiv:2506.15154*, 2025.

A. R. Chow. AI's Influence on Music Is Raising Some Difficult Questions. *TIME*, 2023. URL https://time.com/6340294/ai-transform-music-2023/.

Annie Chu, Patrick O'Reilly, Julia Barnett, and Bryan Pardo. Text2fx: Harnessing clap embeddings for text-guided audio effects. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Eero-Pekka Damskägg, Lauri Juvela, Etienne Thuillier, and Vesa Välimäki. Deep learning for tube amplifier emulation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 471–475. IEEE, 2019.

Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*, 2019.

Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. Musilingo: Bridging music and text with pre-trained language models for music captioning and query response. *arXiv preprint arXiv:2309.08730*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*, 2023.

Chris Donahue, Shih-Lun Wu, Yewon Kim, Dave Carlton, Ryan Miyakawa, and John Thickstun. Hookpad Aria: A copilot for songwriters. In *Extended Abstracts for the Late-Breaking Demo Session of the 25th International Society for Music Information Retrieval Conference*, 2024.

Hao-Wen Dong, Ke Chen, Julian McAuley, and Taylor Berg-Kirkpatrick. Muspy: A toolkit for symbolic music generation. *arXiv preprint arXiv:2008.01951*, 2020.

Qian Dong, Yuchao Lu, Zhiyong Zhang, Xinyu Chen, Hao Zhou, Shuo Zhang, Zhiyuan Zhang, and Lei Yu. Inference-time intervention: Eliciting preference-aligned responses from language models without fine-tuning. *arXiv preprint arXiv:2405.20317*, 2024. URL https://arxiv.org/abs/2405.20317.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain

Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

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.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open, 2024. URL https://arxiv.org/abs/2407.14358.

F. Foscarin, K. Hoedt, V. Praher, A. Flexer, and G. Widmer. Concept-Based Techniques for "Musicologist-Friendly" Explanations in Deep Music Classifiers. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo. Vampnet: Music generation via masked acoustic token modeling. *arXiv preprint arXiv:2307.04686*, 2023.

Josh Gardner, Simon Durand, Daniel Stoller, and Rachel M Bittner. Llark: A multimodal foundation model for music. *arXiv preprint arXiv:2310.07160*, 2023.

Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo. Ace-step: A step towards music generation foundation model. *arXiv preprint arXiv:2506.00045*, 2025.

Matan Gover and Oded Zewi. Music translation: Generating piano arrangements in different playing levels. In *ISMIR*, pages 36–43, 2022.

V. Haunschmid, E. Manilow, and G. Widmer. Towards Musically Meaningful Explanations Using Source Separation. *arXiv preprint arXiv:2009.02051*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020. doi: 10.21105/joss.02154. URL https://doi.org/10.21105/joss.02154. Deezer Research.

Dorien Herremans. Music royalties in the age of ai: Paying artists for ai-generated songs. *WIPO Magazine*, 2025.

Dorien Herremans and Elaine Chew. Morpheus: generating structured music with constrained patterns and tension. *IEEE Transactions on Affective Computing*, 10(4):510–523, 2017.

Dorien Herremans and Abhinaba Roy. Aligning generative music ai with human preferences: Methods and challenges. *arxiv:2511.15038*, 2025.

Dorien Herremans, Ching-Hua Chuan, and Elaine Chew. A functional taxonomy of music generation systems. *ACM Computing Surveys (CSUR)*, 50(5):1–30, 2017.

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 *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

Yichen Huang, Zachary Novack, Koichi Saito, Jiatong Shi, Shinji Watanabe, Yuki Mitsufuji, John Thickstun, and Chris Donahue. Aligning text-to-music evaluation with human preferences, 2025.

Yujia Huang, Adishree Ghatare, Yuanzhe Liu, Ziniu Hu, Qinsheng Zhang, Chandramouli S Sastry, Siddharth Gururani, Sageev Oore, and Yisong Yue. Symbolic music generation with Non-Differentiable rule guided diffusion. February 2024.

Shulei Ji, Jing Luo, and Xinyu Yang. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. *arXiv preprint arXiv:2011.06801*, 2020.

Shulei Ji, Xinyu Yang, and Jing Luo. A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. *ACM Computing Surveys*, 56(1):1–39, 2023.

Faria Binte Kader and Santu Karmaker. A survey on evaluation metrics for music generation. *arXiv preprint arXiv:2509.00051*, 2025.

Jaeyong Kang, Soujanya Poria, and Dorien Herremans. Video2music: Suitable music generation from videos using an affective multimodal transformer model. *Expert Systems with Applications*, 249:123640, 2024.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Féchet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

Sein Kim, Hongseok Kang, Seungyoon Choi, Donghyun Kim, Minchul Yang, and Chanyoung Park. Large language models meet collaborative filtering: An efficient all-round llm-based recommender system. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1395–1406, 2024.

Yonghyun Kim, Wayne Chi, Anastasios Angelopoulos, Wei-Lin Chiang, Koichi Saito, Shinji Watanabe, Yuki Mitsufuji, and Chris Donahue. Music arena: Live evaluation for text-to-music. *arXiv:2507.20900*, 2025.

Peter Knees, Markus Schedl, and Masataka Goto. Intelligent user interfaces for music discovery. *Transactions of International Society for Music Information Retrieval (TISMIR)*, 3(1):165–179, 2020.

Junghyun Koo, Gordon Wichern, François G Germain, Sameer Khurana, and Jonathan Le Roux. Understanding and controlling generative music transformers by probing individual attention heads. 2024.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993, 2023.

Dinh-Viet-Toan Le, Louis Bigo, Mikaela Keller, and Dorien Herremans. Natural language processing methods for symbolic music generation and information retrieval: a survey. *arXiv preprint arXiv:2402.17467*, 2024.

Jinwoo Lee and Kyogu Lee. Do captioning metrics reflect music semantic alignment? *arXiv preprint arXiv:2411.11692*, 2024.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. October 2022.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. MERT: Acoustic music understanding model with Large-Scale self-supervised training. May 2023.

Junling Liu, Chao Liu, Peilin Zhou, Qichen Ye, Dading Chong, Kang Zhou, Yueqi Xie, Yuwei Cao, Shoujin Wang, Chenyu You, et al. Llmrec: Benchmarking large language models on recommendation task. *arXiv preprint arXiv:2308.12241*, 2023.

Renhang Liu, Chia-Yu Hung, Navonil Majumder, Taylor Gautreaux, Amir Ali Bagherzadeh, Chuan Li, Dorien Herremans, and Soujanya Poria. Jam: A tiny flow-based song generator with fine-grained controllability and aesthetic alignment. *arXiv:2507.20880*, 2025.

Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian. Musecoco: Generating symbolic music from text. *arXiv preprint arXiv:2306.00110*, 2023.

Tongyu Lu, Charlotta-Marlena Geist, Jan Melechovsky, Abhinaba Roy, and Dorien Herremans. Melodysim: Measuring melody-aware music similarity for plagiarism detection. *arXiv preprint arXiv:2505.20979*, 2025.

S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017.

Yi Luo and Jianwei Yu. Music source separation with band-split rnn. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1893–1901, 2023. doi: 10.1109/TASLP.2023.3271145.

Yinghao Ma, Anders Øland, Anton Ragni, Bleiz MacSen Del Sette, Charalampos Saitis, Chris Donahue, Chenghua Lin, Christos Plachouras, Emmanouil Benetos, Elio Quinton, et al. Foundation models for music: A survey. *arXiv preprint arXiv:2408.14340*, 2024.

D. Makris, K. R. Agres, and D. Herremans. Generating Lead Sheets with Affect: A Novel Conditional seq2seq Framework. In *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. Muscaps: Generating captions for music audio. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 45–49. IEEE, 2019.

Mark Marrington et al. Composing with the digital audio workstation. *The singer-songwriter handbook*, pages 77–89, 2017.

Andrea Martelloni, Andrew P McPherson, and Mathieu Barthet. Real-time percussive technique recognition and embedding learning for the acoustic guitar. *arXiv preprint arXiv:2307.07426*, 2023.

Marco A. Martínez-Ramírez, Wei-Hsiang Liao, Giorgio Fabbro, Stefan Uhlich, Chihiro Nagashima, and Yuki Mitsufuji. Automatic music mixing with deep learning and out-of-domain data, 2022.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation. *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024a.

Jan Melechovsky, Abhinaba Roy, and Dorien Herremans. Midicaps–a large-scale midi dataset with text captions. *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2024b.

Jan Melechovsky, Ambuj Mehrish, and Dorien Herremans. Sonicmaster: Towards controllable all-in-one music restoration and mastering. *arXiv preprint arXiv:2508.03448*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Adv. Neural Inf. Process. Syst.*, 35:17359–17372, 2022.

David Meredith. Cosiatec and siateccompress: Pattern discovery by geometric compression. In *International society for music information retrieval conference*. International Society for Music Information Retrieval, 2013.

S. Mishra, , B. L. T. Sturm, and S. Dixon. Local Interpretable Model-Agnostic Explanations for Music Content Analysis. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

S. Mishra, E. Benetos, B. L. T. Sturm, and S. Dixon. Reliable Local Explanations for Machine Listening. In *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

Mateusz Modrzejewski and Przemysław Rokita. Text-to-music models and their evaluation methods. *Progress in Polish Artificial Intelligence Research 4, Seria: Monografie Politechniki Łódzkiej Nr. 2437, Wydawnictwo Politechniki Łódzkiej*, 2023.

Alia Morsi, Kana Tatsumi, Akira Maezawa, Takuya Fujishima, and Xavier Serra. Sounds out of pläce? score-independent detection of conspicuous mistakes in piano performances. *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2023.

Jerry Ngo and Yoon Kim. What do language models hear? probing for auditory representations in language models. February 2024.

Truong Q Nguyen. Near-perfect-reconstruction pseudo-qmf banks. *IEEE Transactions on signal processing*, 42(1):65–76, 1994.

Liam Pond and Ichiro Fujinaga. Teaching llms music theory with in-context learning and chain-of-thought prompting: Pedagogical strategies for machines. *arXiv preprint arXiv:2503.22853*, 2025.

Jordi Pons, Xiaoyu Liu, Santiago Pascual, and Joan Serrà. Gass: Generalizing audio source separation with large-scale data. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 546–550. IEEE, 2024.

Isaiah Poritz. Songwriters push for ai protections as congress considers rules. *Bloomberg Law*, Sep 2023. 23 Sept, Available at https://news.bloomberglaw.com/ip-law/songwriters-push-for-ai-protections-as-congress-considers-rules.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The MUSDB18 corpus for music separation, December 2017. URL https://doi.org/10.5281/zenodo.1117372.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proc. of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144, 2016.

Alain Riou, Stefan Lattner, Gaëtan Hadjeres, and Geoffroy Peeters. Pesto: Pitch estimation with self-supervised transposition-equivariant objective. In *International Society for Music Information Retrieval Conference (ISMIR 2023)*, 2023.

Robin San Roman, Pierre Fernandez, Alexandre Défossez, Teddy Furon, Tuan Tran, and Hady Elsahar. Proactive detection of voice cloning with localized watermarking. *arXiv preprint arXiv:2401.17264*, 2024.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096956.

Simon Rouard, Robin San Roman, Yossi Adi, and Axel Roebel. Musicgen-stem: Multi-stem music generation and edition through autoregressive modeling. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Abhinaba Roy, Renhang Liu, Tongyu Lu, and Dorien Herremans. Jamendomaxcaps: A large scale music-caption dataset with imputed metadata. *arXiv preprint arXiv:2502.07461*, 2025a.

Abhinaba Roy, Geeta Puri, and Dorien Herremans. Text2midi-inferalign: Improving symbolic music generation with inference-time alignment. *arXiv preprint arXiv:2505.12669*, 2025b.

P. Samuelson. Generative AI meets copyright. *Science*, 381(6654):158–161, 2023.

Ryosuke Sawata, Naoya Takahashi, Stefan Uhlich, Shusuke Takahashi, and Yuki Mitsufuji. The whole is greater than the sum of its parts: Improving dnn-based music source separation. *arXiv preprint arXiv:2305.07855*, 2023.

Alexander Scarlatos, Yusong Wu, Ian Simon, Adam Roberts, Tim Cooijmans, Natasha Jaques, Cassie Tarakajian, and Cheng-Zhi Anna Huang. Realjam: Real-time human-ai music jamming with reinforcement learning-tuned transformers, 2025.

Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Mo\^ usai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*, 2023.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023.

Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 212–212. IEEE Computer Society, 2024.

Ian Simon, Josh Gardner, Curtis Hawthorne, Ethan Manilow, and Jesse H Engel. Scaling polyphonic transcription with mixtures of monophonic transcriptions. In *ISMIR*, pages 44–51, 2022.

Nikhil Singh, Manuel Cherep, and Pattie Maes. Discovering and steering interpretable concepts in large generative music models. 2025.

Christian J. Steinmetz, Jordi Pons, Santiago Pascual, and Joan Serrà. Automatic multitrack mixing with a differentiable mixing console of neural audio effects. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 71–75, 2021. doi: 10.1109/ICASSP39728.2021.9414364.

Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-U-Net: A Multi-Scale Neural Network for End- to-End Audio Source Separation. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, pages 334–340. ISMIR, November 2018. doi: 10.5281/zenodo.1492417. URL https://doi.org/10.5281/zenodo.1492417.

Kun Su, Judith Yue Li, Qingqing Huang, Dima Kuzmin, Joonseok Lee, Chris Donahue, Fei Sha, Aren Jansen, Yu Wang, Mauro Verzetti, et al. V2meow: Meowing to the visual beat via video-to-music generation. 2024.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

John Thickstun, David Hall, Chris Donahue, and Percy Liang. Anticipatory music transformer. June 2023.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

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.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashvala Vinay and Alexander Lerch. Evaluating generative audio systems and their metrics. *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

Qi Wang, Ruohua Zhou, and Yonghong Yan. Polyphonic piano transcription with a note-based music language model. *Applied Sciences*, 8(3):470, 2018.

Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43, 2023.

Yuyan Wang, Mohit Sharma, Can Xu, Sriraj Badam, Qian Sun, Lee Richardson, Lisa Chung, Ed H Chi, and Minmin Chen. Surrogate for long-term user experience in recommender systems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4100–4109, 2022.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. Muchomusic: Evaluating music understanding in multimodal audio-language models. *arXiv preprint arXiv:2408.01337*, 2024.

Megan Wei, Michael Freeman, Chris Donahue, and Chen Sun. Do music generation models encode music theory? In *International Society for Music Information Retrieval*, 2024.

M. Won, S. Chun, and X. Serra. Toward Interpretable Music Tagging With Self-Attention. *arXiv preprint arXiv:1906.04972*, 2019.

Minz Won, Janne Spijkervet, and Keunwoo Choi. Music classification: beyond supervised learning, towards real-world applications. *arXiv preprint arXiv:2111.11636*, 2021.

Minz Won, Yun-Ning Hung, and Duc Le. A foundation model for music informatics. November 2023.

Minz Won, Yun-Ning Hung, and Duc Le. A foundation model for music informatics. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1226–1230. IEEE, 2024.

Shangda Wu, Dingyao Yu, Xu Tan, and Maosong Sun. Clamp: Contrastive language-music pre-training for cross-modal symbolic music information retrieval. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2023a.

Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J Bryan. Music ControlNet: Multiple time-varying controls for music generation. November 2023b.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023c.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Chenyu Yang, Shuai Wang, Hangting Chen, Wei Tan, Jianwei Yu, and Haizhou Li. Songbloom: Coherent song generation via interleaved autoregressive sketching and diffusion refinement. *arXiv preprint arXiv:2506.07634*, 2025.

Haici Yang, Shivani Firodiya, Nicholas J Bryan, and Minje Kim. Don't separate, learn to remix: End-to-end neural remixing with joint optimization. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 116–120. IEEE, 2022.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, Hao Liu, and Lei Xie. Songeval: A benchmark dataset for song aesthetics evaluation. *arXiv preprint arXiv:2505.10793*, 2025.

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.

Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, et al. Yue: Scaling open foundation models for long-form music generation. *arXiv preprint arXiv:2503.08638*, 2025.

Tian Yun, Zilai Zeng, Kunal Handa, Ashish V Thapliyal, Bo Pang, Ellie Pavlick, and Chen Sun. Emergence of abstract state representations in embodied sequence modeling. November 2023.

Eva Zangerle and Christine Bauer. Evaluating recommender systems: survey and framework. *ACM computing surveys*, 55(8):1–38, 2022.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

Huan Zhang, Vincent KM Cheung, Hayato Nishioka, Simon Dixon, and Shinichi Furuya. Llaqo: Towards a query-based coach in expressive music performance assessment. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Yixiao Zhang, Yukara Ikemiya, Woosung Choi, Naoki Murata, Marco A Martínez-Ramírez, Liwei Lin, Gus Xia, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. Instructmusicgen: Unlocking text-to-music editing for music language models via instruction tuning. *arXiv preprint arXiv:2405.18386*, 2024a.

Yixiao Zhang, Yukara Ikemiya, Gus Xia, Naoki Murata, Marco Martínez, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. Musicmagus: Zero-shot text-to-music editing via diffusion models. *arXiv preprint arXiv:2402.06178*, 2024b.

Daniel Zhao, Daniel Beaglehole, Taylor Berg-Kirkpatrick, Julian McAuley, and Zachary Novack. Steering autoregressive music generation with recursive feature machines. 2025.