# Advanced AI in Music: A Comprehensive Review of Automatic Music Transcription and Generative Technologies

## I. Introduction

### A. The Evolving Landscape of AI in Music

Artificial intelligence is fundamentally reshaping the landscape of music, extending its influence across various facets from intricate analysis and precise transcription to innovative composition and dynamic performance. This transformative evolution is largely propelled by significant advancements in deep learning methodologies, coupled with the increasing availability of sophisticated computational resources and the curation of extensive, meticulously annotated musical datasets. The capability of transcribing music audio into music notation, known as Automatic Music Transcription (AMT), stands as a compelling illustration of human-like intelligence, encompassing perception, cognitive recognition of musical objects, knowledge representation for forming musical structures, and inferential processes for hypothesis testing. This complex task is regarded as a foundational problem within the fields of music signal processing and music information retrieval (MIR).[1] Concurrently, deep learning techniques, particularly Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Unit (GRU) networks, have demonstrated remarkable efficacy in generating new music that closely resembles human-composed works.[3]

The progress in AI for music extends beyond mere technological sophistication; it acts as a powerful catalyst for both musicological research and broader public engagement. By rendering complex acoustic music content accessible in symbolic forms, AI facilitates

large-scale analytical studies of musical structures that were previously laborious and time-consuming. Furthermore, these advancements enable the creation of innovative educational tools and empower both seasoned musicians and aspiring novices in their creative endeavors. This democratization of music creation and analysis underscores AI's profound role as a transformative force, impacting both academic inquiry and practical applications within the musical domain.

### B. Report Objectives: Demystifying AI Music Transcription and Generation

This report aims to provide a comprehensive, detailed, and state-of-the-art overview of AI systems designed for automatic music transcription and music generation. It will delve into the underlying methodologies, architectural choices, and current challenges, offering a clear understanding for technically curious professionals and researchers interested in the intersection of artificial intelligence and music.

### C. Report Structure and Scope

The report is structured into four main sections. Following this introduction, Section II provides a detailed exploration of Automatic Music Transcription (AMT), outlining its foundational concepts, signal processing techniques, deep learning architectures, and challenges. Section III offers a comprehensive review of AI Music Generation, differentiating between symbolic and raw audio approaches, detailing various deep learning models, and discussing inherent challenges. The report concludes in Section IV with a synthesis of key advancements, an exploration of synergies between transcription and generation, and a future outlook for AI in music.

## II. Automatic Music Transcription (AMT): Converting Sound to Symbol

This section thoroughly explains the process of converting acoustic music signals into

symbolic representations, detailing the steps from raw audio to final notation.

## A. Foundational Concepts and Subtasks

### 1. Definition and Analogy to Automatic Speech Recognition (ASR)

Automatic Music Transcription (AMT) is defined as the computational design of algorithms to convert acoustic music signals into a structured form of music notation.[1] This intricate process is widely considered the musical counterpart to Automatic Speech Recognition (ASR), as both tasks involve transforming continuous audio signals into discrete, symbolic representations.[1] However, AMT presents unique complexities. It fundamentally functions as an "inverse problem" in signal processing, where the objective is to reconstruct discrete symbolic musical events from a continuous, complex acoustic mixture. This differs from music synthesis, which is a more direct forward process. AMT must disentangle overlapping frequencies, harmonics, and timbres originating from multiple simultaneous sound sources. This challenge, often referred to as "occlusion," occurs when different notes or instruments share or interfere within the same time-frequency space.[1] Such acoustic interference makes AMT inherently more complex than ASR, where distinct phonemes typically exhibit greater separability. This inherent complexity necessitates the application of highly sophisticated signal processing and advanced machine learning techniques to accurately infer the underlying musical "intent" from the raw audio.

### 2. Core Subtasks: Multi-Pitch Estimation, Onset/Offset Detection, Instrument Recognition, Beat/Rhythm Tracking, Expressive Timing, Score Typesetting

AMT is a multifaceted and challenging endeavor due to the variety of interconnected subtasks it encompasses.[1] A comprehensive AMT system typically integrates several components, each addressing a specific aspect of musical information extraction:

- **Multi-Pitch Estimation (MPE):** This foundational subtask focuses on identifying the number and fundamental frequencies (f0s) of all notes simultaneously present within very short time frames, typically around 10 milliseconds.[1] It is a critical initial step, especially for polyphonic music where multiple notes sound concurrently.[7]
- **Onset and Offset Detection:** This involves precisely identifying the start (onset) and

end (offset) times of individual musical notes.[1] Note onsets are often more readily identifiable due to their percussive nature and distinctive broadband spectrum, and their accurate detection is paramount for overall transcription quality.[8]

- **Instrument Recognition:** This subtask aims to identify the specific instruments playing within the audio recording.[1] It is essential for accurate polyphonic transcription and effective source separation, as the unique timbre (sound quality) of each instrument aids in their differentiation.[4]
- **Beat and Rhythm Tracking:** This component is responsible for determining the underlying pulse, tempo, and rhythmic structure of the music.[1] This temporal information is vital for quantizing note lengths and aligning them to a metrical grid, which is a necessary step for generating standard sheet music notation.[10]
- **Interpretation of Expressive Timing and Dynamics:** This is a particularly challenging subtask that involves capturing the subtle nuances of a musical performance, such as deviations from strict tempo (e.g., *rubato*, *accelerando*, *ritardando*) and variations in loudness (dynamics like *forte*, *piano*, *crescendo*, *decrescendo*).[1] These elements are crucial for rendering a human-like and musically meaningful score.[13]
- **Score Typesetting:** This represents the highest level of transcription, where the extracted musical information is converted into a human-readable musical score. This includes not only individual notes and their precise timing but also various musical symbols for dynamics, articulations, time signatures, and key signatures, presenting a complete notated representation.[1]

AMT is not a monolithic problem but a complex, hierarchical process. Each subtask operates at a distinct level of musical abstraction, building upon the information extracted by the preceding level. For example, accurate Multi-Pitch Estimation at the frame-level is a prerequisite for precise Note-level transcription, which in turn informs Stream-level and Notation-level tasks. This inherent dependency implies that errors or inaccuracies introduced at lower levels can propagate and compound throughout the system, significantly degrading the quality of the final output. Consequently, robust AMT systems necessitate a modular design that can address these subtasks either sequentially or jointly, with a particular emphasis on minimizing errors in the foundational steps.

The following table summarizes these core subtasks:

**Table 1: Key Subtasks in Automatic Music Transcription (AMT)**

| Subtask Name | Description | Key Musical Information Extracted | Challenges |
|---|---|---|---|
| (Multi-)Pitch | Identifying | Fundamental | Overlapping |

| | | | |
|---|---|---|---|
| Estimation | fundamental frequencies (f0s) and pitches of all simultaneous notes. | frequencies, MIDI pitches. | harmonics, polyphony, noise, instrument timbre variations. |
| Onset and Offset Detection | Pinpointing the exact start and end times of individual notes. | Note onset times, note offset times. | Percussive vs. sustained sounds, ambiguous offsets, rapid successions of notes. |
| Instrument Recognition | Identifying the specific instruments playing in the audio. | Instrument timbre, sound source identity. | Mixed signals, similar timbres, unseen instruments, varying recording conditions. |
| Beat and Rhythm Tracking | Determining the underlying pulse, tempo, and rhythmic structure. | Beat times, tempo (BPM), meter, rhythmic patterns. | Expressive timing (rubato), complex syncopation, polyrhythms, varying musical styles. |
| Expressive Timing and Dynamics | Capturing performance nuances like tempo changes and loudness variations. | Rubato, accelerando, ritardando, crescendo, decrescendo, articulation (legato, staccato, accent). | Subjectivity of human performance, subtle continuous variations, lack of precise symbolic equivalents. |
| Score Typesetting | Converting extracted musical data into a human-readable musical score. | Standard notation (notes, rests, clefs, key/time signatures), dynamics, articulations, | Translating continuous performance data to discrete notation, handling complex musical structures, visual |

| | | | |
|---|---|---|---|
| | | phrasing. | layout. |

## B. Audio Signal Processing and Feature Engineering for AMT

### 1. From Raw Audio to Meaningful Representations (Sampling, Bit Depth, Frequency)

Raw audio data, in its unprocessed form, is inherently unstructured and requires significant initial processing to transform it into a format that AI models can effectively analyze and learn from.[16] This preprocessing phase is critical for ensuring the reliability and accuracy of subsequent AI-based analyses. Several fundamental characteristics of audio signals directly influence these initial transformations:

- **Sampling Rate:** This parameter quantifies the number of discrete audio "snapshots" captured per second, typically measured in kilohertz (kHz). A higher sampling rate allows for the capture of more intricate sonic details, but it proportionally increases the resulting file size. For instance, while AI speech recognition systems often employ a 16 kHz sampling rate to balance clarity with computational efficiency, AI-based music generation systems may utilize higher rates, combined with a 24-bit depth, to capture the nuanced variations and richness inherent in musical sounds.[16]
- **Bit Depth:** This property dictates the precision with which each individual audio sample is represented. It directly impacts the dynamic range of the sound, determining the difference between the quietest and loudest sounds that can be captured. Higher bit depths enable the recording of a wider spectrum of sounds with greater fidelity, which is particularly important in music to preserve subtle dynamic shifts and harmonic complexities.[16]
- **Frequency:** Frequency refers to the pitch of a sound, measured in Hertz (Hz). The range of human hearing typically spans from approximately 20 Hz to 20 kHz. During the preprocessing stage, AI models frequently apply filters to remove frequencies deemed irrelevant to the primary task, such as very low or very high frequencies outside the typical human vocal range for speech processing, thereby enhancing the focus on pertinent musical information.[16]

## 2. Preprocessing Techniques: Noise Reduction, Resampling, Segmentation

Preprocessing steps are indispensable for converting raw, unstructured audio data into a clean, usable, and standardized format, making it ready for reliable AI model training and analysis.[16] These techniques include:

- **Data Cleaning:** This involves the removal of unwanted elements from the audio, such as background noise, hums, clicks, or other artifacts that could interfere with accurate musical feature extraction.[16]
- **Resampling:** This technique adjusts the audio's sampling rate to a consistent standard, ensuring uniformity across diverse datasets and compatibility with model input requirements.
- **Normalization:** Normalization standardizes the audio's amplitude levels, preventing louder sections from disproportionately influencing model training and ensuring a balanced dynamic range.
- **Segmentation:** This process involves dividing long audio recordings into smaller, more manageable segments. For music analysis, this often means segmenting the audio into individual notes, phrases, or short temporal windows. Segmentation allows AI models to focus their analysis on specific, relevant parts of the audio without having to process the entire file at once, thereby significantly improving efficiency.[16]

## 3. Time-Frequency Analysis: Short-Time Fourier Transform (STFT) and Spectrograms, Constant Q Transform (CQT)

After initial preprocessing, raw audio is converted into structured data through feature extraction techniques. This step is crucial for machines to effectively process and analyze sound, revealing patterns vital for tasks like music analysis.[16]

- **Short-Time Fourier Transform (STFT) and Spectrograms:** The STFT is a foundational tool in signal processing, particularly effective for analyzing non-stationary signals, which are characteristic of musical audio. It operates by segmenting the continuous audio signal into short, overlapping time windows, or "frames." For each of these frames, the Fourier Transform is computed to determine its frequency spectrum.[16] The resulting collection of these time-localized frequency spectra is then visually represented as a **Spectrogram**.[16] A spectrogram is a two-dimensional graph where the horizontal axis denotes time, the vertical axis represents frequency (ranging from low to high), and the amplitude (loudness or intensity) of a particular frequency component at a given time is indicated by the intensity or color of each point.[16] This visual representation provides an intuitive means of interpreting how the frequency content of a sound evolves over time.[17]

A critical consideration in STFT is the **time-frequency resolution trade-off**: the chosen window size dictates the balance between precise temporal localization (achieved with shorter windows) and accurate frequency representation (achieved with longer windows).[19] Overlapping windows are commonly employed to enhance the continuity and smoothness of the time-frequency representation.[18] Spectrograms are widely applied in music analysis for tasks such as pitch estimation, onset detection, and, fundamentally, music transcription.[19] They can also reveal underlying audio problems like electrical hum, broadband hiss, or transient clicks.[17]

- **Constant Q Transform (CQT):** The CQT is another advanced time-frequency analysis technique that holds particular advantages for music. Unlike the linear frequency spacing of the STFT, CQT employs a constant center frequency-to-resolution ratio. This means its frequency bins are logarithmically spaced, closely mirroring human musical perception, where intervals like octaves are perceived consistently across the frequency spectrum. This characteristic yields a more consistent pattern of sounds with harmonic components in the logarithm-scaled frequency domain, which significantly simplifies the task for AI models to resolve and identify notes played simultaneously, especially in complex polyphonic music.[7] Consequently, CQT coefficients are frequently utilized as robust input features for Convolutional Neural Networks (CNNs) in AMT systems.[7]

The conversion of audio into time-frequency representations, such as spectrograms derived from STFT or CQT, is more than a mere data preprocessing step; it represents a fundamental conceptual shift that bridges the gap between traditional audio signal processing and advanced computer vision techniques. By transforming sound into an "image" where time, frequency, and amplitude are visually represented, AI models, particularly Convolutional Neural Networks (CNNs), can directly apply their powerful image recognition capabilities to identify intricate musical patterns. This enables models to leverage spatial pattern recognition to discern musical features like pitch contours, harmonic structures, and rhythmic patterns, which manifest as visually distinct formations within these representations. This causal link is a cornerstone of the remarkable success observed in modern deep learning applications within Music Information Retrieval (MIR).

## C. Deep Learning Architectures in AMT

The application of deep neural networks, trained on extensive datasets, has driven significant progress in Automatic Music Transcription (AMT).[7] These networks possess the capacity to learn complex representations and intricate musical structures directly from audio data.

## 1. Convolutional Neural Networks (CNNs) for Pitch and Timbre Recognition

Convolutional Neural Networks (CNNs) are extensively utilized in AMT, primarily for processing the two-dimensional time-frequency representations (spectrograms or CQT coefficients) of audio signals.[6] The inherent strength of CNNs lies in their ability to learn hierarchical spatial features, which, in the context of music, translates into recognizing specific pitch patterns, harmonic relationships, and timbral characteristics across both time and frequency axes.[7]

For accurate pitch detection, CNN models can incorporate contextual information by taking a "context window" of frames as input. This allows them to make more informed and robust predictions compared to models that rely solely on frequency features from isolated, single frames.[23] Some advanced approaches employ a two-stage CNN architecture: one CNN is specifically designed to detect note onsets, while a subsequent CNN estimates the probabilities of pitches at those detected onset points.[28] Furthermore, CNN filter kernels can be meticulously designed to operate across different dimensions—such as pitch/pitch class, pitch octaves, and various time scales—to perform sophisticated harmony analysis.[27]

The effectiveness of CNNs in AMT originates from their capacity to function as specialized "feature detectors" for distinct musical "gestures" embedded within the spectrogram. Their convolutional filters, when trained on diverse musical data, implicitly learn to identify the unique frequency components of a note, the characteristic overtone series that defines an instrument's timbre, or the sharp, broadband energy burst associated with a note onset. By tailoring filter shapes and sizes to align with specific musical properties (e.g., vertical filters for recognizing harmonic relationships, horizontal filters for detecting temporal events), CNNs can robustly extract musically relevant information, making them exceptionally effective for the acoustic modeling component of AMT.

## 2. Recurrent Neural Networks (RNNs), LSTMs, and GRUs for Temporal Dynamics

Recurrent Neural Networks (RNNs), particularly their advanced variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Unit (GRU) networks, are exceptionally well-suited for processing the sequential data inherent in music.[3] Their primary advantage lies in their ability to capture and model temporal dependencies and patterns over time, effectively maintaining a "memory" of past inputs.[29] This memory mechanism is crucial for understanding and predicting musical elements such as rhythm, melody, and tonal progression.

**LSTMs** specifically address the vanishing gradient problem, a common challenge in traditional

RNNs that limits their capacity to learn long-term dependencies. They achieve this through an intricate system of internal "gates"—namely, the input gate, forget gate, and output gate—which precisely control the flow of information into and out of a "cell state." This allows LSTMs to selectively retain or discard information over extended sequences, enabling them to remember relevant musical context for prolonged durations.[30]

**GRUs** offer a more streamlined and computationally efficient alternative to LSTMs, also employing gating mechanisms to manage information flow and capture temporal relationships, often with fewer parameters.[3]

These recurrent architectures are frequently employed to model the correlations between pitch combinations over time, forming a critical "music language model" component within comprehensive AMT systems.[15] They are also highly effective in estimating smooth onset probabilities and accurately tracking beats and rhythm, contributing significantly to the temporal precision of transcriptions.[25]

The most effective deep learning architectures for Automatic Music Transcription often employ a hybrid approach, strategically combining the strengths of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs/LSTMs). CNNs excel at extracting robust, localized acoustic features from the time-frequency representations, such as identifying the presence of specific pitches or the sharp attack of an onset. Subsequently, RNNs/LSTMs process these extracted features to model the temporal evolution, dependencies, and sequential context of musical events, such as how individual notes coalesce into melodies, harmonies, and rhythms over time. This synergy enables the system to simultaneously capture both the instantaneous acoustic properties and the overarching musical structure, leading to more coherent, accurate, and musically intelligent transcriptions.

## 3. Transformer Models for Sequence-to-Sequence Transcription

Transformer models, initially developed for Natural Language Processing (NLP) tasks, have rapidly emerged as powerful architectures in AMT due to their innovative self-attention mechanism.[24] This mechanism allows them to effectively capture long-range dependencies and complex contextual relationships within sequences, overcoming some of the inherent limitations of traditional RNNs, such as vanishing gradients.[35] This sequence-to-sequence approach simplifies the transcription process by jointly modeling audio features and language-like output dependencies, thereby reducing the need for highly task-specific architectural designs.[24]

Models like **MT3 (Music Transformer 3)** utilize spectrogram frames as the encoder input, providing a rich acoustic context. An autoregressive decoder then predicts future musical

tokens (e.g., MIDI-like events) based on past outputs, effectively capturing intricate temporal patterns in music, similar to how NLP models process text sequences.[34] Transformers can learn to refer back to previously generated material, making them particularly well-suited for music's inherently repetitive and structured nature.[35]

**Polytune** is another novel Transformer model that directly takes audio inputs and outputs annotated music scores, demonstrating state-of-the-art performance in music error detection.[34]

Transformer models represent a significant paradigm shift in AMT, moving towards more universal and data-driven sequence learning. Their self-attention mechanism enables them to model global dependencies across entire musical sequences, capturing complex musical grammar and structure directly from tokenized representations of music (e.g., MIDI-like events) without the vanishing gradient issues often encountered with RNNs. This development suggests a future where highly flexible, multi-task models, potentially pre-trained on vast and diverse audio and symbolic datasets, could handle a wide array of Music Information Retrieval (MIR) tasks with reduced reliance on specialized architectural engineering, simplifying the development process and potentially improving generalization across different musical contexts.

The following table provides a comparison of these core deep learning architectures in the context of AMT:

**Table 2: Comparison of Core Deep Learning Architectures for AMT**

| Architecture | Primary Role in AMT | Strengths | Limitations/Challenges | Key SOTA Models/Concepts |
|---|---|---|---|---|
| **Convolutional Neural Networks (CNNs)** | Feature extraction from time-frequency representations (spectrograms, CQT) for pitch, onset, timbre recognition. | Excellent at local feature detection, spatial pattern recognition; robust to variations. | Limited in modeling long-term temporal dependencies without recurrent layers. | CREPE, components of Onsets and Frames, CNNs for CQT analysis.[6] |

| | | | | |
|---|---|---|---|---|
| **Recurrent Neural Networks (RNNs) / LSTMs / GRUs** | Modeling temporal dependencies, rhythm tracking, note duration, and musical language modeling. | Capable of capturing sequential patterns and long-term dependencies (LSTMs/GRUs); effective for onset/offset detection. | Traditional RNNs suffer from vanishing/exploding gradients; LSTMs/GRUs are computationally intensive and slower for very long sequences. | Onsets and Frames (with LSTMs), various beat tracking and music language models.[3] |
| **Transformer Models** | Sequence-to-sequence transcription, capturing long-range dependencies and global musical structure. | Self-attention allows modeling very long-range dependencies; highly parallelizable; can simplify complex architectures. | Computationally expensive for very long sequences (though improvements exist); requires large datasets for effective training. | MT3, Polytune, generic encoder-decoder Transformers for MIDI-like output.[24] |

## D. Addressing Polyphony and Instrument Separation

### 1. Challenges of Overlapping Frequencies and Timbre

Polyphonic music, characterized by the simultaneous sounding of multiple notes or instruments, presents one of the most formidable challenges for Automatic Music Transcription (AMT).[1] This inherent complexity primarily arises from the intricate interaction and overlap of harmonics originating from different sound sources within the acoustic signal.[7] Musical objects, such as individual notes, frequently occupy the same time-frequency regions within a spectrogram, leading to an "occlusion" problem that makes it exceedingly difficult to

differentiate between individual notes or instruments.[1] For instance, the harmonic series of notes played concurrently within a chord can either destructively interfere, resulting in a decrease or complete elimination of certain frequencies, or constructively interfere, leading to an increase in overall amplitude. Both scenarios significantly complicate accurate multi-pitch estimation.[7] Furthermore, the diverse timbres (the unique sound qualities) of different instruments pose a substantial issue. The model must distinguish between sources that may share similar fundamental frequencies but possess distinct overtone structures, which define their characteristic sound. The precise synchronization of onsets and offsets between different voices, a common feature in ensemble music, also violates the typical assumption of statistical independence between sources, an assumption that would otherwise simplify the separation process.[1]

## 2. Multi-Task Learning and Source Separation Integration (e.g., Cerberus)

To effectively mitigate the challenges posed by polyphonic music, **audio source separation** is often integrated as a crucial preliminary or concurrent step within AMT systems.[4] This process involves the estimation and inference of individual source signals from a mixed acoustic observation.[1] State-of-the-Art (SOTA) approaches increasingly integrate source separation directly into the deep learning network architecture, frequently employing

**multi-task deep learning** techniques.[4] This integrated approach allows the model to learn shared representations that mutually benefit both the separation and transcription tasks.

A prominent example of this integrated methodology is **Cerberus**, a novel deep learning architecture designed to simultaneously separate an audio recording of a musical mixture into its constituent single-instrument recordings and transcribe these instruments into a human-readable format.[6] Cerberus extends existing source separation networks by incorporating a dedicated "head" for transcription. By training each head with different loss functions, the model jointly learns how to separate and transcribe multiple instruments, demonstrating that these two tasks are highly complementary. This joint learning paradigm results in networks that exhibit superior performance in both separation and transcription, and also generalize more effectively to unseen musical mixtures.[39] Other models, such as

**Open-Unmix** and **Band-Split Recurrent Neural Networks (BSRNN)**, are also utilized for accurate vocal and instrument separation, particularly in complex scenarios like choral music, where multiple voices blend intricately.[38]

The performance ceiling for Automatic Music Transcription, particularly in polyphonic and multi-instrument scenarios, is not merely dependent on incremental improvements in note detection algorithms but is intrinsically linked to the system's ability to effectively perform

source separation. Treating these as entirely separate, sequential tasks (first separate, then transcribe) introduces potential error propagation and limits overall accuracy. The prevailing trend towards multi-task learning, as exemplified by models like Cerberus, demonstrates that a shared internal representation of musical information significantly benefits both separation and transcription simultaneously. This suggests that future breakthroughs in AMT for complex music will likely originate from models that deeply integrate source separation capabilities, recognizing the fundamental interdependence of disentangling sound sources and accurately identifying their musical content.

## E. State-of-the-Art (SOTA) Models and Advanced Considerations

### 1. Key Models: Onsets and Frames, CREPE, Basic-Pitch

Significant progress in Automatic Music Transcription (AMT) has been driven by the development of sophisticated deep learning models:

- **Onsets and Frames (OaF):** Developed by Google Magenta, OaF is a seminal model for automatic polyphonic piano music transcription that has achieved state-of-the-art results.[6] Its core innovation lies in decomposing the complex task of note detection into two distinct neural network stacks: one specifically trained to detect *onset frames* (the very beginning of each note) and another trained to detect *every frame where a note is active* (the "frames" or sustain portion).[8] The output from the onset detector is utilized in two ways: it is fed as an additional input to the frame detector, and it also constrains the final output, ensuring that new notes are only initiated when the onset detector indicates high confidence.[8] This dual-objective approach substantially improved accuracy by incentivizing the model to predict note beginnings precisely, thereby preserving crucial musical events.[9] Recent variants have explored architectural modifications, such as replacing recurrent layers with Convolutional Neural Networks (CNNs) and incorporating Context-Aware Modules (CAM) to efficiently capture temporal vicinity.[10]
- **CREPE:** This deep learning model is highly regarded for its fundamental frequency (f0) estimation capabilities. It processes input audio signals by converting them into spectrograms and subsequently feeding these representations through Convolutional Neural Networks (CNNs).[6] CREPE's robust performance makes it a frequent and effective component in many pitch detection pipelines.
- **Basic-Pitch:** This model is frequently referenced in the context of Multi-Pitch Estimation (MPE) benchmarks, consistently demonstrating competitive F1 scores across various

datasets, indicating its strong performance in identifying multiple simultaneous pitches.[5]

- **Transformer-based models:** Beyond hybrid CNN-RNN architectures, generic encoder-decoder Transformers are increasingly demonstrating performance equivalent to custom deep neural networks for transcription. This simplifies the model design process by jointly modeling audio features and language-like output dependencies.[24]
  **Polytune** is a novel Transformer model that directly accepts audio inputs and produces annotated music scores, showcasing state-of-the-art performance in music error detection.[34]
  **MT3 (Music Transformer 3)** also leverages spectrograms as encoder input for token-based output, effectively capturing temporal patterns in music.[34]

## 2. Modeling Expressive Performance: Rubato, Accelerando, Dynamics, Articulation

Accurately transcribing the expressive nuances of a musical performance, such as subtle variations in tempo and dynamics, remains a significant and complex challenge for AMT systems.[1] Human performers deliberately shape these parameters to convey musical meaning, emotion, and structural understanding within a composition.[14] Key expressive elements that AMT systems strive to capture include:

- **Rubato:** An Italian term meaning "stolen time," it refers to a flexible disregard of certain notated rhythmic and tempo properties for the sake of expressive performance. This can manifest as a solo melody moving with subtle rhythmic redistribution against a steady accompaniment, or as the entire musical texture accelerating or slowing down, often with an eventual return to the original tempo.[40]
- **Accelerando (accel.):** A gradual increase in the musical tempo.[12]
- **Ritardando (rit.):** A gradual decrease in the musical tempo.[12]
- **Dynamics:** Indications of loudness, typically conveyed using Italian terms such as *forte* (loud), *piano* (quiet), *crescendo* (gradually louder), and *decrescendo* (gradually quieter).[12]
- **Articulation:** This refers to the connection or separation between notes, as well as the accent level at the beginning of a note (its attack). Examples include *legato* (playing or singing smoothly and connected), *staccato* (playing or singing notes more separated), and *accent* (playing or singing a note with extra stress or emphasis).[12]

State-of-the-art models are beginning to address these complexities. Research indicates that features like note duration, pitch, metrical strength, phrase position, and overall tempo are crucial for predicting expressive transformations.[43] Deep learning models are being developed to analyze and generate expressive piano performances, aiming to capture vivid micro-timing, rich polyphonic texture, varied dynamics, and sustain pedal effects.[44] Furthermore,

multimodal networks like

**VioPose** estimate precise 4D human pose data to analyze subtle movements such as vibrato, demonstrating the intricate link between physical performance and musical expression. This detailed understanding of performance gestures is vital for achieving comprehensive and human-like transcriptions.[45]

A significant challenge in Automatic Music Transcription lies in bridging the "semantic gap" between the acoustic reality of a performance and the symbolic representation in traditional music notation. While AMT can accurately detect fundamental musical attributes like pitches, onsets, and offsets, capturing the full richness of human expressive performance—the subtle, continuous variations in timing (rubato, accelerando) and dynamics, as well as precise articulation—is considerably more complex. Standard symbolic formats like MIDI and conventional sheet music often represent these nuances as abstract textual or graphical cues rather than precise, continuous data. This implies that achieving notation-level AMT requires not just sophisticated acoustic modeling but also an advanced "music language model" [15] capable of inferring performer intent and translating continuous acoustic expressive features into appropriate discrete symbolic markings. This moves the task beyond simply identifying "what notes were played" to accurately conveying "how they were played," which remains a key frontier in achieving truly human-like transcriptions.

## 3. Current Limitations and Open Research Problems

Despite remarkable progress, Automatic Music Transcription (AMT) continues to face several significant open challenges, particularly concerning polyphonic music and generalization.[1] Common issues observed in current AMT outputs include:

- **Octave and Semitone Errors:** Misidentification of notes by an octave or a semitone, often due to harmonic ambiguities or the complex interplay of overtones in polyphonic signals.
- **Missed Notes:** Failure to detect notes, especially within dense chords where overlapping harmonics can obscure individual pitches.[1]
- **Extra Notes:** Spurious note detections, often manifesting as harmonic errors, particularly in the presence of unseen timbres or complex acoustic environments.[1]
- **Merged or Split Notes:** Incorrect segmentation of notes, where a single sustained note might be transcribed as multiple shorter notes, or multiple distinct notes might be incorrectly merged into one longer note.[1]

Broader challenges for the field include:

- **Data Scarcity and Diversity:** A persistent lack of general and sizable annotated

datasets, particularly for multi-instrument mixtures and diverse musical genres beyond piano music.[5] This limitation significantly hinders the ability of supervised models to generalize effectively to real-world music recordings with varied instrumentation and styles.[46]

- **Generalization to Unseen Timbre and Styles:** Current models often struggle to accurately transcribe music featuring instruments or playing styles not adequately represented in their training data.[1] This indicates a need for more robust feature representations or domain adaptation techniques.
- **Accurate Interpretation of Expressive Timing and Dynamics:** As previously discussed, fully capturing the subtle nuances of human expressive performance remains a complex task. The continuous and subjective nature of elements like rubato, accelerando, and dynamic variations makes their precise symbolic representation challenging.
- **Computational Demands:** Training and deploying state-of-the-art deep learning models for AMT can be computationally intensive, requiring significant hardware resources and processing time.
- **Real-time Performance:** Achieving high-accuracy AMT in real-time for complex polyphonic music remains a challenge, limiting applications in live performance analysis or interactive music systems.

## F. Output Formats: MIDI, Piano Rolls, and Sheet Music Generation

The output of an AMT system can take several forms, each serving different purposes in music applications:

- **MIDI (Musical Instrument Digital Interface):** This is a widely used symbolic representation of music. MIDI files contain event-based data, specifying note onsets, offsets, pitch (MIDI note number), velocity (loudness), and other control changes.[7] MIDI is highly editable and flexible, allowing musicians to change instruments, rearrange melodies, or adjust tempos in digital audio workstations (DAWs) or notation software.[48] It is also highly compact, leading to lower energy consumption for training, generation, and storage compared to raw audio.[49]
- **Piano Rolls:** A piano-roll representation is a visual display of pitches over time, often used as an intermediate or final output for AMT systems.[1] It graphically shows notes as horizontal bars, with the vertical axis representing pitch and the horizontal axis representing time. This format is intuitive for visualizing note activity, duration, and polyphony.
- **Sheet Music (Music Notation):** This is the highest level of symbolic representation, aiming to produce a human-readable musical score. This involves not only the notes, their pitches, and durations but also rhythmic structures, time signatures, key signatures,

dynamics, articulations, and other musical symbols.[1] While MIDI can be converted to sheet music by software, achieving a musically intelligent and aesthetically pleasing score from complex audio remains a significant challenge, often requiring sophisticated post-processing for timing quantization and voice separation.[1] Commercial software like Sibelius's AudioScore and ScoreCloud offer audio-to-score conversion capabilities.[1]

# III. AI Music Generation: Creating New Compositions

This section explores how AI systems create new music, detailing the fundamental methodologies, deep learning architectures, and current challenges in generating compositions across various formats.

## A. Fundamental Methodologies: Symbolic vs. Raw Audio Generation

The field of AI music generation broadly divides into two main methodological approaches, each with distinct advantages and disadvantages concerning data representation, control, and realism:

### 1. Symbolic Generation (MIDI, Sheet Music): Advantages and Disadvantages

Symbolic music generation utilizes AI technologies to create symbolic representations of music, such as MIDI files, sheet music, or piano rolls.[51] The core of this approach lies in training models to learn the underlying structures of music, including chord progressions, melodies, and rhythmic patterns, to generate compositions with logical and structured musical coherence.[51] These models typically handle discrete note data, and the generated results can be directly played back through synthesizers or further converted into audio.[51]

**Advantages:**

- **Structured and Controllable:** Symbolic representations inherently capture musical structure (notes, rhythm, harmony), making it easier to control specific musical elements like pitch, duration, and velocity.[51] Composers can enforce chord progressions or time signatures.[55]

- **Editability and Flexibility:** MIDI files are highly editable in Digital Audio Workstations (DAWs) or notation software, allowing for easy manipulation of instruments, melodies, and tempos.[48]
- **Computational Efficiency:** Symbolic models, especially those using MIDI, operate on much smaller datasets and representations compared to raw audio. This results in significantly lower computational resource requirements and faster training and generation times, contributing to a lower environmental impact.[49]
- **Copyright Clarity:** Generating MIDI files can offer clearer licensing terms for commercial usage, as they are not direct copies of existing audio recordings.[49]
- **Accessibility:** Allows non-musicians to create music without needing traditional instrumental skills.[48]

**Disadvantages:**

- **Limited Expressiveness:** Symbolic representations often struggle to capture the subtle nuances and richness of human expressive performance, such as intricate timbral variations, micro-timing (rubato), or complex articulations that are inherent in raw audio.[51] The resulting playback can sound "mechanical" or "cold" without additional human intervention or sophisticated rendering.[44]
- **Requires Synthesis:** Symbolic outputs (MIDI, sheet music) are not directly audible; they require a separate synthesis step (e.g., a software synthesizer or virtual instrument) to be converted into sound.[51]
- **Generalization Challenges:** While good for structured music, generalizing to highly diverse or unconventional musical styles can be challenging.[60]

## 2. Raw Audio Generation (Waveforms, Spectrograms): Advantages and Disadvantages

Raw audio music generation directly produces the audio signal of music, including waveforms or spectrograms, which can be played back immediately without further processing.[51] This approach operates closer to the recording and mixing stages of music production, capable of producing music content with complex timbres and a high degree of realism.[51]

**Advantages:**

- **High Fidelity and Realism:** Raw audio models can generate highly realistic and nuanced sounds, capturing intricate timbral details, expressive variations, and subtle sonic textures that are difficult to represent symbolically.[51]
- **Direct Playback:** The output is an immediately audible sound file, eliminating the need for a separate synthesis step.[51]

- **Comprehensive Information:** Raw audio preserves all information from the original sound wave, including aspects like vibrato, trills, and pitch bending, which MIDI tokenizations may fail to represent accurately.[34]

**Disadvantages:**

- **Computational Intensity:** Generating raw audio, especially high-fidelity waveforms, typically requires substantial computational resources (e.g., powerful GPUs) and significant training time due to the large number of samples in audio format.[51]
- **Challenges with Long-Term Coherence:** While excelling at local sound quality, raw audio generation models often face difficulties in controlling the overall structure and musical logic over extended durations, potentially leading to compositions that lack long-term coherence or form.[51]
- **Limited Granular Control:** Directly manipulating specific musical elements (e.g., changing a single note's pitch or duration) in a raw audio waveform is significantly more complex than in symbolic formats.[51]
- **Data Scarcity for High-Quality Training:** Acquiring large datasets of high-quality, labeled raw audio for training can be challenging.
- **Copyright Ambiguity:** The legal status of AI-generated raw audio, particularly concerning training data derived from existing copyrighted music, remains a complex and debated issue.[49]

The following table provides a comparative analysis of symbolic versus raw audio music generation methodologies:

**Table 3: Symbolic vs. Raw Audio Music Generation: A Comparative Analysis**

| Feature | Symbolic Music Generation | Raw Audio Music Generation |
|---|---|---|
| **Data Representation** | MIDI files, sheet music, piano rolls (discrete notes, events). | Waveforms, spectrograms (continuous audio signals). |
| **Output** | Editable symbolic files; requires synthesizer for sound. | Directly playable audio files. |
| **Control & Structure** | High granular control over notes, rhythm, harmony; excels at structured compositions. | Limited direct granular control; challenges in maintaining long-term musical structure. |

| | | |
|---|---|---|
| **Expressiveness & Fidelity** | Can lack subtle nuances of human performance; often sounds "mechanical" without further processing. | High fidelity and realism; captures intricate timbral details and expressive variations. |
| **Computational Cost** | Relatively low; smaller file sizes, faster training/generation. | High; requires significant computational resources and time. |
| **Primary Use Cases** | Composition, education, prototyping, interactive music, game audio, score generation. | Sound design, realistic soundscapes, vocal synthesis, high-quality track generation. |
| **Key Architectures** | RNNs (LSTMs, GRUs), Transformers (e.g., Music Transformer, MuseNet). | WaveNet, Jukebox, Diffusion Models (e.g., MusicLM, MeLoDy). |
| **Advantages** | Editability, efficiency, clear structure, accessibility, lower environmental impact. | High realism, direct audibility, captures full sonic detail. |
| **Disadvantages** | Limited expressiveness, requires synthesis, can sound artificial. | High computational cost, challenges with long-term coherence, complex editing, copyright issues. |

## B. Deep Learning Architectures for Music Generation

Deep learning models have revolutionized music generation, moving beyond rule-based systems to create complex and coherent compositions. Various architectures are employed, each leveraging different strengths to model musical patterns and structures.

### 1. Recurrent Neural Networks (RNNs), LSTMs, and GRUs for Sequential

### Composition

Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks and Gated Recurrent Unit (GRU) networks, are fundamental to sequential music composition due to their inherent ability to process and generate sequences of data.[3] Music, being a time-dependent sequence of notes, melodies, and rhythms, naturally benefits from RNNs' capacity to maintain a "memory" of past inputs and use this context to predict future elements.[29]

**LSTMs** are particularly adept at capturing long-term dependencies within musical sequences, a critical aspect for generating coherent melodies and harmonies that extend beyond a few notes.[32] They achieve this through a sophisticated gating mechanism—comprising input, forget, and output gates—which regulates the flow of information into and out of a "cell state." This allows LSTMs to selectively remember or forget musical patterns over long durations, enabling them to learn complex structures like recurring chord progressions or motifs.[30]

**GRUs** offer a similar capability with a simpler architecture, often providing comparable performance with reduced computational overhead.[3]

When applied to music generation, LSTMs learn musical patterns by modeling the temporal dependencies of note sequences and optimizing their parameters through backpropagation to fit the probability distribution of training data.[32] The dynamic adjustment of gating signals and the long-term transfer of the cell state enable these networks to learn from historical musical data and generate new sequences that reflect learned patterns. For instance, DeepBach utilizes LSTMs to generate Bach-style harmonies, producing harmonically consistent chord progressions.[51] While effective for consistent melodies and rhythms, LSTMs can still struggle with maintaining long-term coherence across entire movements or songs, as very long sequences can still present challenges for their memory mechanisms.[51]

## 2. Generative Adversarial Networks (GANs) for Realistic Outputs

Generative Adversarial Networks (GANs) have emerged as powerful deep learning techniques for generating realistic synthetic data, including music.[67] A GAN architecture consists of two competing neural networks: a

**generator** and a **discriminator**.[67]

The **generator** network's role is to create new data samples (e.g., musical sequences,

spectrograms) that closely resemble real data. It takes a random noise vector as input and transforms it into a musical output, aiming to deceive the discriminator.[67] The

**discriminator** network, on the other hand, acts as a binary classifier. It receives both real music samples from the training dataset and fake music samples generated by the generator, and its objective is to correctly classify the input as either real or fake.[67]

This adversarial training process involves a continuous, iterative "game" between the two networks. The generator strives to produce increasingly realistic music to fool the discriminator, while the discriminator simultaneously improves its ability to identify fake data.[67] This competitive dynamic drives both networks to improve, ultimately leading the generator to produce music samples that are virtually indistinguishable from real human-composed music.[67] GANs learn the underlying patterns and structures of a given musical dataset, allowing them to generate new music that is consistent in style and structure with the training data.[68] Variants like Deep Convolutional Generative Adversarial Networks (DCGANs) are particularly effective at capturing and understanding these underlying patterns.[68] While GANs excel at generating high-quality, realistic outputs, they can sometimes be challenging to train due to their adversarial nature and may require careful parameter tuning.

### 3. Variational Autoencoders (VAEs) for Latent Space Exploration

Variational Autoencoders (VAEs) offer another approach to music generation by combining the principles of autoencoders with random sample generation.[32] VAEs learn low-dimensional, continuous representations (latent spaces) of music from large datasets. They achieve this by encoding input music into a compressed latent vector and then decoding this vector back into a musical output, aiming to reconstruct the original input. Crucially, VAEs introduce a probabilistic component, ensuring that the latent space is well-structured and continuous, allowing for smooth interpolation between different musical pieces and the generation of novel variations by sampling from this latent space.[32]

The theoretical foundations of VAEs, like other deep learning models, involve the **backpropagation algorithm** for gradient calculation and parameter updates during training, with the goal of minimizing the difference between generated and real music.[32] Appropriate

**loss functions**, such as Mean Squared Error (MSE) and Cross-Entropy, are defined to measure this discrepancy.[32] The training objective for VAEs in music generation is to maximize the diversity and creativity of the generated music while maintaining consistency in its structure and tonality.[32] Models like MusicVAE utilize a hierarchical latent vector model to learn long-term structure, first outputting embeddings for sub-sequences and then generating

each subsequence from these embeddings.[54] VAEs are particularly useful for exploring musical variations and generating compositions with a consistent style.

### 4. Transformer Models for Long-Term Coherence (e.g., Music Transformer, MuseNet, MusicGen)

Transformer models, with their powerful self-attention mechanisms, have become dominant in music generation, particularly for their ability to capture long-term coherence and complex musical structures.[51] Unlike RNNs that process data sequentially, Transformers parallelize computation across all elements in an input sequence, significantly improving training speed and scalability.[37]

The core of a Transformer's success in music generation lies in its **attention mechanism**.[35] This dynamic and highly parallelizable mechanism allows the model to weigh the importance of different parts of the input sequence when predicting the next output. For music sequences exhibiting repetition, a Transformer can learn to "attend" to specific material that appeared much earlier in the piece, enabling it to refer back to and develop previously generated elements.[35] This capacity to model dependencies over very long ranges addresses a significant limitation of recurrent neural networks.[35]

- **Music Transformer:** This model demonstrates the first successful use of Transformers in generating music that exhibits long-term structure.[70] It employs a modified relative attention mechanism that allows it to generate minute-long compositions with compelling internal consistency and to coherently elaborate on given motifs.[70] By representing music as a sequence of discrete tokens, Music Transformer takes a language-modeling approach to generative music.[70]
- **MuseNet:** Developed by OpenAI, MuseNet is a deep neural network capable of generating 4-minute musical compositions, blending diverse musical styles and utilizing multiple instruments.[72] It learns and predicts patterns of harmony, rhythm, and style by analyzing vast quantities of MIDI files, anticipating the next note in each sequence. MuseNet employs a general unsupervised learning technique similar to GPT-2, which predicts the next token in a sequence.[72]
- **MusicGen:** This is a generative music Transformer model that uses an EnCodec to create discrete audio tokens, an autoregressive Transformer to predict the next token, and an EnCodec decoder to output an audio signal.[74] MusicGen's self-attention heads learn to understand and represent diverse musical elements, from instrument recognition to more complex musical attributes, enabling nuanced music generation.[74]

These Transformer-based models excel at capturing long-term dependencies and generating

intricate melodies, demonstrating their ability to produce structurally sound and musically coherent compositions over extended durations.[71]

### 5. Diffusion Models for High-Fidelity Audio Synthesis (e.g., MusicLM, MeLoDy)

Diffusion models represent a cutting-edge class of generative models that have achieved state-of-the-art performance in generating high-fidelity audio, including music.[61] The fundamental principle behind diffusion models involves a multi-step reverse inference process: they gradually recover a clean audio signal from random noise.[71] During generation, the model iteratively denoises a noisy input, progressively refining it until a high-quality, stable audio output is achieved.[61] This process allows them to capture fine details in timbre and tonal variations, making them exceptionally suitable for generating realistic and diverse audio content.[71]

- **MusicLM:** Introduced by Google, MusicLM is a pioneering model that generates high-fidelity music from text descriptions, such as "a calming violin melody backed by a distorted guitar riff".[63] MusicLM frames conditional music generation as a hierarchical sequence-to-sequence modeling task, capable of generating music at 24 kHz that remains consistent over several minutes.[63] It employs a hierarchy of three Language Models (LMs) for semantic, coarse acoustic, and fine acoustic modeling, allowing it to understand and generate music based on diverse text prompts related to genre, instruments, tempo, scenarios, or subjective feelings.[63] MusicLM has demonstrated superior performance in both audio quality and adherence to text descriptions, and can even transform whistled or hummed melodies according to a specified text style.[75]
- **MeLoDy (Music, LM, Diffusion):** Developed to address the computational expense of MusicLM, MeLoDy is an LM-guided diffusion model that generates state-of-the-art quality music audio while significantly reducing the number of forward passes compared to MusicLM (e.g., 95.7% to 99.6% reduction for 10s to 30s music sampling).[63] MeLoDy inherits the highest-level LM from MusicLM for semantic modeling, which determines the overall arrangement of melody, rhythm, dynamics, timbre, and tempo.[63] It then employs a novel dual-path diffusion (DPD) model and an audio VAE-GAN to efficiently decode the conditioning semantic tokens into waveforms.[63] The DPD model is designed to simultaneously model coarse and fine acoustics by incorporating semantic information into segments of latents via cross-attention at each denoising step.[63] MeLoDy's advancements lie in its practical advantages like sampling speed and infinitely continuable generation, alongside its high musicality, audio quality, and text correlation.[63]

Diffusion models excel in capturing fine details in timbre and tonal variations, making them particularly effective for generating high-quality and diverse audio content.[71] While MusicLM

and MeLoDy leverage semantic modeling for broader musical structure, the explicit mechanisms for ensuring long-term coherence across extended generated music pieces are a continuous area of refinement within diffusion models, often involving hierarchical generation or conditioning strategies.[54]

The following table provides an overview of state-of-the-art AI music generation models, categorized by their underlying deep learning architectures:

**Table 4: Overview of State-of-the-Art AI Music Generation Models**

| Architecture | Key Architectures/ Concepts | Primary Strengths | Limitations | Output Format |
|---|---|---|---|---|
| **Recurrent Neural Networks (RNNs) / LSTMs / GRUs** | DeepBach, Performance_ RNN | Excels at capturing temporal dependencies, consistent melodies and rhythms; effective for sequential data. | Struggles with very long-term coherence (vanishing gradients); can sound mechanical without expressive rendering. | MIDI, Symbolic (piano rolls, sheet music). |
| **Generative Adversarial Networks (GANs)** | MuseGAN, DCGAN, ProGAN | Generates highly realistic outputs; learns complex patterns from data distribution. | Can be challenging to train (mode collapse); less direct control over specific musical elements. | Spectrograms, MIDI, Raw Audio (via vocoder). |
| **Variational Autoencoders (VAEs)** | MusicVAE | Learns low-dimension al latent representation s; good for style transfer and interpolation; | Can struggle with high-fidelity detail; outputs may lack sharpness compared to GANs/Diffusio | MIDI, Symbolic. |

| | | | | |
|---|---|---|---|---|
| | | enables exploration of musical variations. | n. | |
| **Transformer Models** | Music Transformer, MuseNet, MusicGen, Pop Music Transformer | Excellent at long-term coherence and global structure via self-attention; highly parallelizable. | Computationally intensive for very long sequences; requires large datasets. | MIDI, Symbolic (tokens), Raw Audio (via vocoder/decoder). |
| **Diffusion Models** | MusicLM, MeLoDy, AudioGen, MusicGen | Generates high-fidelity, realistic audio; excels at capturing subtle timbral details; robust generation process. | Can be slow for inference (though MeLoDy addresses this); challenges in explicit long-term structural control. | Raw Audio (waveforms, spectrograms). |

## C. Challenges and State-of-the-Art in Music Generation

Despite the impressive advancements, AI music generation still faces several significant challenges that researchers are actively addressing:

### 1. Ensuring Long-Term Coherence and Musical Structure

One of the most persistent challenges in AI music generation is maintaining long-term coherence and a meaningful musical structure across extended compositions.[3] While models

can generate musically plausible short segments or phrases, ensuring that these segments integrate into a cohesive, aesthetically pleasing, and structurally sound full-length piece remains difficult. Music relies heavily on repetition, variation, and development of motifs and themes over various timescales, from phrases to entire movements.[70] Traditional RNNs often struggle with this due to vanishing gradients, leading to compositions that become random or drift off-topic over time.[60] Transformers, with their self-attention mechanism, have made significant strides in this area by allowing the model to "look back" at distant parts of the sequence, enabling the generation of minute-long compositions with compelling structure and coherent elaborations on given motifs.[70] Diffusion models like MusicLM also address this by casting generation as a hierarchical sequence-to-sequence modeling task, leveraging semantic LMs to guide overall arrangement.[63] Hierarchical generation approaches, where high-level structures are composed first, followed by low-level details, are also being explored to ensure whole-song coherence.[54]

### 2. Granular Control and Expressiveness in Generated Music

While AI can generate technically correct music, imbuing it with human-like expressiveness and allowing granular control over specific musical parameters remains a complex task.[51] Human musicians introduce subtle variations in timing (rubato, accelerando), dynamics, and articulation to convey emotion and interpretation.[14] AI-generated compositions can sometimes sound "generic" or "over-processed," lacking the nuanced emotional depth and unique touch that human artists bring.[56] Models are being developed to generate expressive piano performances by capturing vivid micro-timing, rich polyphonic texture, and varied dynamics.[44] Furthermore, text-to-music models like MusicLM aim to allow control over key, BPM, and other characteristics through text prompts, moving towards more granular creative control.[62] However, translating abstract textual descriptions into precise expressive musical parameters is still an active research area.

### 3. Computational Demands and Data Scarcity

State-of-the-art music generation models, particularly those producing high-fidelity raw audio (e.g., Jukebox, diffusion models), demand substantial computational resources for both training and inference.[51] This can be a barrier to entry for many researchers and artists. While models like MeLoDy are designed to reduce computational passes significantly compared to MusicLM, the overall resource requirement remains high.[63]

Additionally, the availability of large-scale, high-quality, and diverse datasets for training remains a challenge. While some datasets like MAESTRO and MAPS exist for piano [22], comprehensive datasets for multi-instrumental or multi-genre music with detailed annotations (e.g., expressive markings, instrument separation) are scarcer.[5] This limits the generalizability and diversity of generated outputs.

### 4. Ethical and Copyright Implications

The rise of AI music generation introduces complex ethical and legal questions, particularly regarding copyright and ownership.[49] Questions arise about who owns the rights to music created by an AI: the developer of the AI, the user who prompts it, or a combination? The use of existing copyrighted music for training AI models also raises concerns about potential infringement.[58] While some AI music tools offer clear licensing terms for commercial use of their generated output, the broader legal landscape is still evolving.[49] There are also concerns about the potential reduction in demand for human musicians and the homogenization of music if many creators rely on similar AI tools, leading to a saturation of similar-sounding compositions.[56] Balancing AI's technical innovations with the emotional depth of human artistry is crucial for the sustainable development of this field.[58]

## D. Applications and Future Directions in Music Creation

AI music generation offers a wide array of applications and promises exciting future directions:

- **Automated Composition and Accompaniment:** AI can assist composers by generating ideas, creating backing tracks, or even completing songs, significantly speeding up the creative process.[1] It can also provide automatic accompaniment for live performers.
- **Personalized Music Recommendations:** AI can generate music tailored to individual user preferences, moods, or specific scenarios.[3]
- **Sound Design and Production:** AI tools can manipulate sound effects, instruments, and synthesizers with ease, creating atmospheric textures or specific sound effects, and assisting in mixing and mastering processes.[58]
- **Music Education:** AI generators can make music production more accessible to aspiring artists without formal training, allowing them to bring their ideas to life.[56]
- **Interactive Experiences:** AI can generate adaptive music that responds in real-time to user input or listener mood, creating dynamic and engaging experiences.[80]

- **Exploration of New Musical Styles:** AI can create unique sounds and combinations that humans might not conceive of, pushing the boundaries of musical genres and fostering experimentation.[79]

The future of AI in music will likely involve a deeper integration of symbolic and raw audio generation techniques, combining structural coherence with high-fidelity realism. Further research will focus on improving granular control, addressing ethical considerations, and developing models that can truly capture and generate the full spectrum of human musical expression.

# IV. Conclusion and Future Outlook

## A. Synthesis of Key Advancements

The landscape of AI in music has undergone a profound transformation, marked by significant advancements in both Automatic Music Transcription (AMT) and music generation. AMT has progressed from a challenging signal processing problem to a sophisticated application of deep learning, capable of converting complex acoustic signals into symbolic representations. This involves a hierarchical process, starting from precise multi-pitch estimation and onset/offset detection, moving through instrument recognition and rhythm tracking, and culminating in the challenging task of interpreting expressive performance nuances and typesetting full musical scores. The conversion of audio into time-frequency representations, such as spectrograms and Constant Q Transforms (CQT), has been pivotal, enabling Convolutional Neural Networks (CNNs) to act as powerful feature detectors for musical "gestures." The integration of Recurrent Neural Networks (RNNs), particularly LSTMs and GRUs, has allowed models to capture the temporal dynamics and long-term dependencies essential for understanding musical flow. More recently, Transformer models have emerged as universal sequence learners, capable of modeling global musical structures and language-like dependencies with remarkable efficiency. The progression in AMT for polyphonic music has also underscored the fundamental interdependence of source separation and transcription, leading to multi-task learning architectures that jointly address these challenges.

In parallel, AI music generation has evolved from rule-based systems to highly creative deep learning models. The field is broadly divided into symbolic generation (MIDI, sheet music), which offers high structural control and computational efficiency, and raw audio generation (waveforms, spectrograms), which excels in producing high-fidelity and realistic sonic

textures. Architectures like LSTMs and GRUs have proven effective for sequential composition, while Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have pushed the boundaries of realistic and diverse output generation. Transformer models, such as Music Transformer and MuseNet, have demonstrated unprecedented capabilities in maintaining long-term coherence and complex musical structures. The latest frontier, diffusion models like MusicLM and MeLoDy, are achieving state-of-the-art audio fidelity, bridging the gap between semantic understanding and realistic sound synthesis.

The following table summarizes the major challenges and current approaches in AI music, encompassing both transcription and generation:

**Table 5: Major Challenges and Current Approaches in AI Music**

| Challenge Area | Description | Current Approaches/Solutions | Impact on Field |
|---|---|---|---|
| **Polyphony & Source Separation** | Overlapping frequencies and timbres from multiple instruments make individual note/instrument identification difficult. | Multi-task learning, integrated source separation (e.g., Cerberus, Open-Unmix), shared musical representations. | Enables more accurate transcription of complex music; improves generalization across mixtures. |
| **Expressive Performance** | Capturing subtle, continuous human nuances (rubato, dynamics, articulation) beyond discrete notes. | Advanced feature engineering (e.g., note duration, metrical strength), specialized deep learning architectures (e.g., VioPose), multi-modal learning. | Moves AMT towards human-like scores; enhances realism and musicality in generated music. |
| **Long-Term Coherence & Structure** | Maintaining musical logic and form over extended compositions | Transformer models (self-attention), hierarchical LMs | Essential for generating full, coherent pieces; improves structural |

| | | | |
|---|---|---|---|
| | (minutes-long). | (MusicLM), cascaded diffusion models, explicit structural conditioning. | integrity of compositions. |
| **Data Scarcity & Generalization** | Lack of diverse, large-scale annotated datasets for various instruments/genres; models struggle with unseen data. | Semi-supervised learning, pre-training on large unlabelled datasets, domain adaptation techniques, synthetic data generation. | Improves model robustness and applicability to real-world scenarios. |
| **Computational Demands** | Training and inference of SOTA models require significant hardware and time. | Model optimization (e.g., MeLoDy's reduced passes), efficient architectures, distributed computing. | Increases accessibility and practical deployment; enables real-time applications. |
| **Granular Control** | Difficulty in precisely controlling specific musical elements in generated output. | Text-to-music conditioning, disentangled latent spaces (VAEs), attention mechanisms for specific attributes. | Empowers artists with more creative agency; allows for targeted modifications. |
| **Ethical & Copyright Issues** | Ambiguity of ownership for AI-generated music; concerns over training data usage and human displacement. | Development of clear licensing frameworks, emphasis on AI as a creative assistant, responsible data sourcing. | Shapes the legal and economic landscape; fosters collaboration between AI and human artists. |

## B. Synergies Between Transcription and Generation

The fields of Automatic Music Transcription and AI music generation, while distinct, share profound synergies that are increasingly being leveraged. Advances in AMT directly benefit music generation by providing high-quality, structured symbolic data (MIDI, piano rolls) from real-world performances. These transcribed datasets serve as invaluable training material for generative models, enabling them to learn complex musical patterns, styles, and even expressive nuances from human performances. Conversely, the development of sophisticated generative models, particularly those capable of high-fidelity audio synthesis, can also inform and improve AMT. For instance, generative models can be used to create diverse synthetic datasets for training AMT systems, especially for challenging scenarios like polyphony or rare instruments, thereby mitigating data scarcity. Furthermore, the "music language models" developed for generation, which capture the statistical regularities and structural rules of music, can be integrated into AMT systems to improve the musical plausibility and coherence of transcriptions, acting as a form of post-processing or prior knowledge. This symbiotic relationship accelerates progress across the entire spectrum of AI in music.

## C. The Future of AI in Music: Research Frontiers and Societal Impact

The future of AI in music is poised for continued rapid advancement, with several key research frontiers emerging. One critical area involves the development of truly end-to-end models that can seamlessly integrate transcription, generation, and even performance aspects, moving beyond modular pipelines. This would entail models capable of perceiving, understanding, creating, and adapting music in a holistic manner. Achieving more sophisticated control over expressive parameters, such as rubato and articulation, in both transcription and generation remains a significant challenge, requiring deeper understanding and modeling of human musical intent. Research into multimodal AI, combining audio, symbolic, and even visual (e.g., performer gestures) data, will likely lead to more nuanced and comprehensive music intelligence systems.

From a societal perspective, AI in music promises to democratize music creation and education, making sophisticated tools accessible to a broader audience. It can unlock new forms of creative expression for artists, serving as an intelligent assistant that handles tedious tasks or provides novel ideas. However, careful consideration of ethical implications, including copyright, fair compensation for human artists, and the preservation of human artistic unique expression, will be paramount. The ongoing dialogue between AI researchers, musicians, and

legal experts will be crucial in shaping a future where AI serves as an empowering force, enhancing human creativity rather than diminishing it, ultimately enriching the global musical landscape.

**Works cited**


1. Automatic Music Transcription: An Overview - University Lab Sites, accessed on July 21, 2025, https://labsites.rochester.edu/air/publications/benetatos19automaticmusic.pdf
2. Automatic Music Transcription - An Overview - Spotify Research, accessed on July 21, 2025, https://research.atspotify.com/publications/automatic-music-transcription-an-overview
3. (PDF) Music Generation using RNN-LSTM with GRU - ResearchGate, accessed on July 21, 2025, https://www.researchgate.net/publication/375671182_Music_Generation_using_RNN-LSTM_with_GRU
4. A Survey on Automatic Music Transcription - IRE Journals, accessed on July 21, 2025, https://www.irejournals.com/formatedpaper/1704431.pdf
5. Investigating an Overfitting and Degeneration Phenomenon in Self-Supervised Multi-Pitch Estimation - arXiv, accessed on July 21, 2025, https://arxiv.org/html/2506.23371v1
6. A Comparison of Deep Learning Methods for Timbre Analysis in Polyphonic Automatic Music Transcription - MDPI, accessed on July 21, 2025, https://www.mdpi.com/2079-9292/10/7/810
7. Automatic Music Transcription using Convolutional Neural Networks and Constant-Q transform - arXiv, accessed on July 21, 2025, https://arxiv.org/html/2505.04451
8. ONSETS AND FRAMES: DUAL-OBJECTIVE PIANO TRANSCRIPTION - ISMIR, accessed on July 21, 2025, https://archives.ismir.net/ismir2018/paper/000019.pdf
9. Onsets and Frames: Dual-Objective Piano Transcription, accessed on July 21, 2025, https://magenta.tensorflow.org/onsets-frames
10. Machine Learning Techniques in Automatic Music Transcription: A Systematic Survey - arXiv, accessed on July 21, 2025, https://arxiv.org/html/2406.15249v1
11. A Comprehensive Review on Music Transcription - MDPI, accessed on July 21, 2025, https://www.mdpi.com/2076-3417/13/21/11882
12. Other Aspects of Notation – Open Music Theory - VIVA's Pressbooks, accessed on July 21, 2025, https://viva.pressbooks.pub/openmusictheory/chapter/other-aspects-of-notation/
13. Score and Performance Features for Rendering Expressive Music Performances - Music Encoding Initiative, accessed on July 21, 2025, https://music-encoding.org/conference/abstracts/abstracts_mec2019/Dasaem%20Jeong%20Music%20Encoding%20Conference%202019.pdf
14. Computational Models of Expressive Music Performance: A Comprehensive and

Critical Review - Frontiers, accessed on July 21, 2025, https://www.frontiersin.org/journals/digital-humanities/articles/10.3389/fdigh.2018.00025/full
15. END-TO-END AUTOMATIC MUSIC TRANSCRIPTION OF POLYPHONIC QANUN AND OUD MUSIC USING DEEP NEURAL NETWORK | Request PDF - ResearchGate, accessed on July 21, 2025, https://www.researchgate.net/publication/384452531_END-TO-END_AUTOMATIC_MUSIC_TRANSCRIPTION_OF_POLYPHONIC_QANUN_AND_OUD_MUSIC_USING_DEEP_NEURAL_NETWORK
16. Getting Started with Audio Data: Processing Techniques and Key ..., accessed on July 21, 2025, https://medium.com/@zilliz_learn/getting-started-with-audio-data-processing-techniques-and-key-challenges-420dc5233163
17. Understanding spectrograms - iZotope, accessed on July 21, 2025, https://www.izotope.com/en/learn/understanding-spectrograms
18. Short-time Fourier transform (STFT) | Advanced Signal Processing Class Notes - Fiveable, accessed on July 21, 2025, https://library.fiveable.me/advanced-signal-processing/unit-6/short-time-fourier-transform-stft/study-guide/DIDcR67AFZ58LkOX
19. Short-time Fourier transform (STFT) | Advanced Signal Processing Class Notes | Fiveable, accessed on July 21, 2025, https://library.fiveable.me/advanced-signal-processing/unit-1/short-time-fourier-transform-stft/study-guide/OprLM7X5xQrmPIAp
20. spectrogram - Chrome Music Lab, accessed on July 21, 2025, https://musiclab.chromeexperiments.com/spectrogram/
21. Spectrogram - Wikipedia, accessed on July 21, 2025, https://en.wikipedia.org/wiki/Spectrogram
22. Music Transcription Using Deep Learning - CS229, accessed on July 21, 2025, https://cs229.stanford.edu/proj2017/final-reports/5242716.pdf
23. Piano Music Transcription Using Convolutional Neural Networks - CS230 Deep Learning - Stanford University, accessed on July 21, 2025, http://cs230.stanford.edu/projects_winter_2019/reports/15808060.pdf
24. SEQUENCE-TO-SEQUENCE PIANO TRANSCRIPTION ... - ISMIR, accessed on July 21, 2025, https://archives.ismir.net/ismir2021/paper/000030.pdf
25. CNN architecture for onset detection. - ResearchGate, accessed on July 21, 2025, https://www.researchgate.net/figure/CNN-architecture-for-onset-detection_fig3_319978212
26. Deep Learning Methods for Instrument Separation and Recognition - QMRO Home, accessed on July 21, 2025, https://qmro.qmul.ac.uk/xmlui/handle/123456789/92140
27. MODELING MUSIC MODALITY WITH A KEY-CLASS INVARIANT PITCH CHROMA CNN - ISMIR, accessed on July 21, 2025, https://archives.ismir.net/ismir2019/paper/000065.pdf
28. A Two-Stage Approach to Note-Level Transcription of a Specific Piano - MDPI, accessed on July 21, 2025, https://www.mdpi.com/2076-3417/7/9/901

29. DeepBeats: Music Genre Classification using LSTM and RNN - ijrpr, accessed on July 21, 2025, https://ijrpr.com/uploads/V5ISSUE4/IJRPR25922.pdf
30. Tutorial on RNN | LSTM: With Implementation - Analytics Vidhya, accessed on July 21, 2025, https://www.analyticsvidhya.com/blog/2022/01/tutorial-on-rnn-lstm-gru-with-implementation/
31. An Efficient Hidden Markov Model with Periodic Recurrent Neural Network Observer for Music Beat Tracking - MDPI, accessed on July 21, 2025, https://www.mdpi.com/2079-9292/11/24/4186
32. LSTM-Based Music Generation Technologies - MDPI, accessed on July 21, 2025, https://www.mdpi.com/2073-431X/14/6/229
33. Transformer-Based Music Language Modelling and Transcription | Request PDF, accessed on July 21, 2025, https://www.researchgate.net/publication/363426384_Transformer-Based_Music_Language_Modelling_and_Transcription
34. Detecting Music Performance Errors with Transformers - arXiv, accessed on July 21, 2025, https://arxiv.org/html/2501.02030v1
35. Tradformer: A Transformer Model of Traditional Music ... - IJCAI, accessed on July 21, 2025, https://www.ijcai.org/proceedings/2022/0681.pdf
36. Transformer-Based Approaches for Automatic Music Transcription - Papers With Code, accessed on July 21, 2025, https://paperswithcode.com/paper/transformer-based-approaches-for-automatic
37. Transformer-based Note level Automatic Drum-Set Transcription, accessed on July 21, 2025, https://www.ewadirect.com/proceedings/ace/article/view/18302/pdf
38. DLVS4Audio2Sheet: Deep learning-based vocal separation for audio into music sheet conversion - InK@SMU.edu.sg, accessed on July 21, 2025, https://ink.library.smu.edu.sg/context/sis_research/article/10163/viewcontent/4._DLVS4Audio2Sheet_Deep_Learning_based_Vocal_Separation_for_Audio_into_Music_Sheet_Conversion.pdf
39. [1910.12621] Simultaneous Separation and Transcription of Mixtures with Multiple Polyphonic and Percussive Instruments - arXiv, accessed on July 21, 2025, https://arxiv.org/abs/1910.12621
40. The Uses of Rubato in Music, Eighteenth to Twentieth Centuries - Scholarship @ Claremont, accessed on July 21, 2025, https://scholarship.claremont.edu/context/ppr/article/1123/viewcontent/RosenblumSpring1994.exe.pdf
41. Rubato: ritardando and accelerando - MuseScore, accessed on July 21, 2025, https://musescore.org/en/node/9383
42. (PDF) Automatic Characterization of Dynamics and Articulation of Expressive Monophonic Recordings - ResearchGate, accessed on July 21, 2025, https://www.researchgate.net/publication/242370288_Automatic_Characterization_of_Dynamics_and_Articulation_of_Expressive_Monophonic_Recordings
43. A Machine Learning Approach to Discover Rules for Expressive Performance Actions in Jazz Guitar Music - PMC, accessed on July 21, 2025,

https://pmc.ncbi.nlm.nih.gov/articles/PMC5167744/
44. Expressive MIDI-format Piano Performance Generation - arXiv, accessed on July 21, 2025, https://arxiv.org/html/2408.00900v1
45. arXiv:2411.13607v2 [cs.CV] 25 Nov 2024, accessed on July 21, 2025, https://arxiv.org/pdf/2411.13607?
46. Daily Papers - Hugging Face, accessed on July 21, 2025, https://huggingface.co/papers?q=automatic%20music%20transcription
47. A Deep Learning-Based Piano Music Notation Recognition Method - PMC, accessed on July 21, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC9184194/
48. AI MIDI Converter: Automatic Music Transcription & Audio to MIDI - Music Demixer, accessed on July 21, 2025, https://freemusicdemixer.com/midi
49. Ethics of AI MIDI, accessed on July 21, 2025, https://midi.org/ethics-of-ai-midi
50. [2504.18502] Music Tempo Estimation on Solo Instrumental Performance - arXiv, accessed on July 21, 2025, https://arxiv.org/abs/2504.18502
51. Applications and Advances of Artificial Intelligence in Music Generation:A Review - arXiv, accessed on July 21, 2025, https://arxiv.org/html/2409.03715v1
52. Generating Music using AI - Lund University Publications, accessed on July 21, 2025, https://lup.lub.lu.se/student-papers/record/9093922/file/9093927.pdf
53. Music Generation Using Deep Learning and Generative AI: A Systematic Review, accessed on July 21, 2025, https://www.researchgate.net/publication/388213469_Music_Generation_Using_Deep_Learning_and_Generative_AI_A_Systematic_Review
54. tatsuropfgt/papers: read paper memo - GitHub, accessed on July 21, 2025, https://github.com/tatsuropfgt/papers
55. MusicLang - Hugging Face, accessed on July 21, 2025, https://huggingface.co/musiclang
56. 12 AI Music Generators That Create Original Songs in 2025 | DigitalOcean, accessed on July 21, 2025, https://www.digitalocean.com/resources/articles/ai-music-generators
57. Conditioning Deep Generative Raw Audio Models for Structured Automatic Music - arXiv, accessed on July 21, 2025, https://arxiv.org/abs/1806.09905
58. The Pros and Cons of AI Music Production - Lalals, accessed on July 21, 2025, https://lalals.com/blog/blog-pros-and-cons-ai-music-production
59. Using AI in music: the impact for the music creator - Telefónica, accessed on July 21, 2025, https://www.telefonica.com/en/communication-room/blog/ai-music-impact-music-creator/
60. A Review of AI Music Generation Models, Datasets, and Evaluation Techniques Milind Uttam Nemade1, accessed on July 21, 2025, https://spast.org/techrep/article/download/5262/537/10498
61. From Discrete Tokens to High-Fidelity Audio Using Multi-Band Diffusion - OpenReview, accessed on July 21, 2025, https://openreview.net/forum?id=dOanKg3jKS
62. Lyria - Google DeepMind, accessed on July 21, 2025, https://deepmind.google/models/lyria/

63. NeurIPS Poster Efficient Neural Music Generation - NeurIPS 2025, accessed on July 21, 2025, https://neurips.cc/virtual/2023/poster/71058
64. Mo\^usai: Text-to-Music Generation with Long-Context Latent Diffusion - ResearchGate, accessed on July 21, 2025, https://www.researchgate.net/publication/367529681_Mousai_Text-to-Music_Generation_with_Long-Context_Latent_Diffusion
65. [2506.00045] ACE-Step: A Step Towards Music Generation Foundation Model - arXiv, accessed on July 21, 2025, https://arxiv.org/abs/2506.00045
66. DeepClassic: Music Generation with Neural Neural Networks - Stanford University, accessed on July 21, 2025, https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1204/reports/custom/report11.pdf
67. How Generative Adversarial Networks (GANs) Work - Keymakr, accessed on July 21, 2025, https://keymakr.com/blog/how-generative-adversarial-networks-gans-work/
68. Multi-Genre Symbolic Music Generation using Deep Convolutional Generative Adversarial Network - ITM Web of Conferences, accessed on July 21, 2025, https://www.itm-conferences.org/articles/itmconf/pdf/2023/03/itmconf_icdsia2023_02002.pdf
69. [2504.02586] Deep learning for music generation. Four approaches and their comparative evaluation - arXiv, accessed on July 21, 2025, https://arxiv.org/abs/2504.02586
70. MUSIC TRANSFORMER: GENERATING MUSIC WITH LONG-TERM STRUCTURE - OpenReview, accessed on July 21, 2025, https://openreview.net/pdf?id=rJe4ShAcF7
71. Features, Models, and Applications of Deep Learning in Music Composition - Science Publishing Group, accessed on July 21, 2025, https://article.sciencepublishinggroup.com/pdf/10.11648.j.ajist.20250903.11
72. Musenet : Generates 4-minute musical compositions combining diverse musical styles and instruments. - AI Tools, accessed on July 21, 2025, https://app.aibase.com/en/details/10267
73. Musenet : Music Generation using Abstractive and Generative Methods - ResearchGate, accessed on July 21, 2025, https://www.researchgate.net/publication/363856706_Musenet_Music_Generation_using_Abstractive_and_Generative_Methods
74. understanding and controlling generative music transformers by probing individual attention heads - Mitsubishi Electric Research Laboratories, accessed on July 21, 2025, https://www.merl.com/publications/docs/TR2024-032.pdf
75. Daily Papers - Hugging Face, accessed on July 21, 2025, https://huggingface.co/papers?q=MusicLM
76. MusicLM: Generating Music From Text | Request PDF - ResearchGate, accessed on July 21, 2025, https://www.researchgate.net/publication/367461777_MusicLM_Generating_Music_From_Text
77. Efficient Neural Music Generation - OpenReview, accessed on July 21, 2025,

https://openreview.net/forum?id=cxazQGSsQa¬eId=MrHGkNn0pH
78. The Pros and Cons of AI in the Music Industry - Slime Green Beats, accessed on July 21, 2025, https://slimegreenbeats.com/blogs/music/the-pros-and-cons-of-ai-in-the-music-industry
79. The Pros and Cons of Using AI in Music Production - SOUNDRAW Blog, accessed on July 21, 2025, https://soundraw.io/blog/post/pros-and-cons-of-using-ai-in-music-production
80. AI Music Generation Models: The Future of Sound and the Role of Meta's AudioCraft, accessed on July 21, 2025, https://www.appypiedesign.ai/blog/ai-music-generation-models