

Artificial Intelligence in Music Generation: Models, Evaluation, Applications, and Future Prospects

Tiffany Chiu

*Jericho Senior High School, New York, USA
tiffanychiu3000@gmail.com*

Abstract. Music creation through the use of artificial intelligence (AI) is an emerging and rapidly developing field. This paper presents a comprehensive review of the current state of AI music generation, covering the historical development of computer-assisted music production and AI-assisted music from early analog and digital tools to modern neural network architectures, and highlighting key developments such as MIDI, DAWs, plugins, and early algorithmic composition systems. It also examines symbolic and audio-based music representations, including MIDI, sheet music, waveforms, and spectrograms, and evaluates generative models such as GANs, LSTMs, Transformers, VAEs, and diffusion models, analyzing their various capabilities and limitations. Applications in areas such as content creation, gaming, healthcare, and marketing also demonstrate AI's growing global impacts. This review also compares subjective, objective, and combined evaluation strategies used to assess new AI music models and addresses challenges and potential problematic areas in current studies and research. Finally, future research directions are discussed, including improved generative techniques, interdisciplinary integration, and real-time interactive systems, suggesting pathways for researchers to enhance creativity, expressiveness, and practical application in AI-assisted music production.

Keywords: Artificial Intelligence, Music Generation, Music Evaluation, Generative Models, Music Representation

1. Introduction

The recent rapid development of new artificial intelligence (AI) technologies has enabled the automatic generation of complex melodies, harmonies, and arrangements, greatly changing the way music is composed. Advances in neural networks and generative models have expanded the capabilities of AI music systems, and new hybrid approaches combining various methods of generation are being explored, in order to produce high-quality computer-generated music.

Research on music production dates back decades, from live performance-focused methods and early analog recordings to the increasing prevalence of computational music tools. For example, the introduction of digital synthesizers and plugins offered music producers previously unprecedented control and creative flexibility in their work. In the modern day, a large portion of music production is done completely digitally.

By building on this foundation, AI has become an important force in contemporary music production. Modern AI systems, through the use of architectures such as Generative Adversarial Networks (GANs), Transformers, and diffusion models and the analyzing of a analyzing large-scale datasets, can generate new, complex musical pieces. Today, the use of AI allows people to experiment with new genres in music, change compositions in real time, and explore new forms of musical expression, establishing AI as an critical component of modern music-making.

Lately, AI music generation is changing from being experimental to becoming a useful tool for a variety of uses. Its current applications span a wide range of fields, including healthcare, marketing, and gaming, and enable processes such as personalized music therapy and adaptive, immersive, soundtracks.

Despite significant progress, several challenges remain about the creation of music through AI. Generating unique and interesting music is still very difficult, and methods of evaluation often lack standardization, which makes it hard to consistently assess and compare new and current tools. As the applications of AI-generated music expand, there is a growing demand for high-quality and faster, real-time outputs. Addressing these challenges requires further analysis of current research and future developments to advance technical capabilities and push current boundaries and limitations.

In this paper, I aim to provide a systematic review of the latest advancements in AI music generation. It looks at the potential and limitations of current methods and the systems used to evaluate them. And it also presents various challenges and proposes new directions for future exploration, to inspire further innovation and the continued development and use of AI as a core tool in music creation and production.

2. History and related work

The history of computer music generation spans over six decades. Early systems relied on rule-based systems and probabilistic approaches. The introduction and rise of deep learning marked a turning point, driving a period of rapid advancement and widespread experimentation in computation music. Contemporary approaches now are able to create symbolic processing, and also can extend beyond to the direct generation of high-fidelity audio, supporting tasks that range from reproducing traditional acoustic instruments to inventing entirely novel sounds.

Recently, AI music generation programs have made significant advancements, especially in generation output quality. In particular, the development of Generative Adversarial Networks (GANs), Transformers, and various diffusion models have helped to enhance variety, structural depth, and expressiveness in generated outputs. Additionally, emerging hybrid frameworks that merge symbolic and audio-based generation are helping to strengthen both the formal organization and musical richness of machine-produced music. These advancements have expanded the boundaries of algorithmic composition and created better ways to explore and directions for innovation in both art and technology.

2.1. Early music production

In the early 1900s, the recording of music was done on analog machinery, mixing boards and tape-based recording systems. In the 1970s, the introduction of synthesizers allowed musicians to generate sounds and design a wide range of timbres and effects through the modulation of waveforms such as sine and triangle waves. This innovation significantly broadened the creative technologies available to composers and producers at the time [1].

2.2. Digital Audio Workstations (DAWs) and plugins

In the late 1900s, as advanced digital music technology quickly expanded, Digital Audio Workstations (DAWs) began to be used. DAW platforms consolidated tasks such as recording, editing, mixing, and composing into a single software system, making the overall workflow of music creation streamlined and accessible. DAWs allow producers to compose within a single software environment, simplifying technical processes and lowering entry barriers for music production [2-4]. The growing adoption of DAWs also gave rise to an extensive ecosystem of plugins, software extensions that added to or modified the core functionalities of DAWs. As a result, they introduced a large range of effects and tools, further expanding creative possibilities available to producers.

2.3. Artificial intelligence

Advancements in digital technology have steadily brought AI into the field of music creation and production. By processing large-scale music datasets, AI systems are able to identify structural patterns, extract key features, and generate original compositions. One early integration of AI into music was Max/MSP. One of the first practical uses of AI-driven methods in music production., Max/MSP was an interactive environment for audio programming that enabled musicians to design custom instruments and sound effects through code. As these technologies evolved, new machine learning tools were introduced, including those that could compose music from curated datasets and automate technical processes such as mixing and mastering. Now, contemporary AI systems can go even further, being able to both replicate established musical styles and create entirely novel musical forms, resulting in an expansion of the creative possibilities available to composers, producers, and artists.

2.4. Modern music production

In the present day, the majority of music production is almost entirely digital at every stage, from composition and arrangement to mixing, mastering, and sound editing. The constantly expanding range of plugins, ranging from resonators to advanced convolution reverbs, offers producers virtually limitless possibilities for sound design and experimentation. The integration of AI technologies has pushed these capabilities even further, enabling the creation of brand-new pieces from scratch. AI-powered tools have significantly extended the range of creative resources that are available to musicians, facilitating the development of new and unique genres, approaches, and soundscapes.

3. Methodology

3.1. Music representation

AI music generation can be separated broadly into symbolic and audio-based approaches [5]. Symbolic generation produces abstract representations (such as MIDI files, sheet music, or piano roles), operating on discrete data and emphasizing the modeling of musical elements such as melody, harmony, rhythm, and structure. This approach allows for interpretable and controllable composition but is limited in timbre and expressive nuance.

On the other hand, audio music generation directly synthesizes sound signals, including waveforms or spectrograms by working with continuous audio data. It focuses on realism, timbre,

and expressive detail. However, while it can produce highly realistic music, it often requires substantial computational resources and may face challenges in maintaining long-term musical structure.

How music data is represented is an important aspect of AI music generation systems, because it influences the quality of generated music and how models interpret input data and produce outputs. Different methods of representing music are able to capture various features of music, such as structure, timbre, or rhythm. And each has its advantages and disadvantages. The following outlines several widely adopted representation approaches and the contexts in which they are typically applied (Table 1 & 2).

Table 1. Symbolic music generation representation

Sheet Music	Sheet music remains the classical method of representing music, using a music staff and various symbolic markings for features such as pitch, rhythm, dynamics, and expression. Within the context of AI music generation, this representation is also valuable for producing outputs that remain interpretable and consistent with traditional music theory. Models can be created to both transcribe existing pieces into sheet music or generate entirely new composition scores.
Musical Instrument Digital Interface (MIDI)	MIDI is a standardized digital system that encodes musical information (such as notes, pitch, dynamics, and tempo) as “performance instructions” to be rendered by hardware synthesizers or music software, making it a flexible tool for AI-driven music generation. In symbolic music generation, MIDI is the most commonly used representation format, often serving as both input and output because of its near-universal compatibility and ability to precisely handle basic levels of musical detail [4].

Table 2. Audio music generation representation

Audio Waveforms	Audio waveforms capture audio directly in the time-domain, making it the most faithful but also the most data-intensive representation; an example is WaveNet, a deep neural network capable of generating realistic speech and music in the form of raw audio waveforms [6]. While this format retains every detail of the original generated signals and creates very detailed audio information, processing them requires significant computational power and memory.
Spectrogram	A spectrogram represents audio in the frequency domain. They illustrate how different frequency components change over time. Common variants such as Short-Time Fourier Transform (STFT), Mel, and Constant-Q spectrograms are widely used for music tasks. Because of their ability to represent both audio's frequency structure and temporal characteristics, spectrograms are important in music analysis, classification, and generation. For example, the model Tacotron 2 takes text-based prompts and outputs raw spectrograms, which are then converted into waveforms [7].

3.2. Classification

These two broad categories (audio and symbolic) of music generation can be further divided into various types of generative models, each suited to specific tasks with unique strengths, data types, and musical goals. Following a survey conducted by Chen et al., 2024, some major types of these models are outlined below (Table 3).

Table 3. Classifications of music generation models

Diffusion Models	Diffusion models create realistic music by gradually refining noisy signals into clear audios. Neural networks are trained by gradually adding, or “diffusing”, random noise to data, and then reversing that process to generate quality audio. Recent developments include Moûsai, a cascading latent diffusion model capable of generating high-quality stereo music from text prompts [8]. Noise2Music is another series of diffusion models able to generate music clips from text through the use of both a generator and cascader diffusion model [9].
Generative Adversarial Networks (GANs)	GANs create realistic and high-quality music by training a generator and discriminator in opposition, making them well-suited for producing complex and varied audio. GAN models have all made notable achievements in generating individual and layered, multi-track compositions. These models include WaveGAN, an application of GANs to unsupervised raw audio creation [10]; and MuseGAN, a combination of three models for multi-track music generation capable of generating music from scratch, without human input [11].
Long Short-Term Memory Networks (LSTMs)	LSTMs are highly effective at processing sequential data and capturing long-term dependencies capable of producing musically coherent and expressive pieces. For example, LSTM models BachBot [12] and DeepBach [13] were used to create polyphonic music modeled after the chorales of Johann Sebastian Bach. However, these models have significant drawbacks in that they require the use of large amounts of training data and high computational demands, which can restrict their use in settings where resources are limited.
Transformers	Transformers use mechanisms in self-attention in order to process sequences more efficiently, making them strong at modeling long-term relationships and complex patterns. MusicLM [14] combined Transformer-based models with audio synthesis to create music generation from a text input through hierarchical sequence-to-sequence modeling. However, as with LSTMs, Transformers are limited by their demand for substantial computational resources. The Music Transformer generated structured segments created by capturing recurring musical motifs over time to create human-like compositions with a reduced, linear sequence-length memory requirement to utilize less memory resources compared to other similar models [15].
Variational Autoencoders (VAEs)	VAEs learn compressed latent spaces that can be sampled to generate new music. For example, MIDI-VAE enables style transfer between different pieces by modifying features such as pitch, instrumentation, and dynamics (through note duration and velocity), resulting in new, adapted melodies or brand-new mixtures of entire songs [16]. On the other hand, the Conditional VAE allows more conditional information to be introduced. Compared to GANs or Transformers, however, VAE-generated music may be less coherent or sound less musical.
Other Models	Beyond the main approaches, researchers have also explored techniques such as Convolutional Neural Networks (CNNs), a deep learning model that uses convolutional layers to automatically detect and learn hierarchical features from structured data; types of Recurrent Neural Networks (RNNs), a model designed for sequential data where each step’s output depends on both the current input and information carried from previous steps; and methods of combining multiple different models to leverage various strengths and improve overall performance.

4. Evaluation methods

As AI-driven music generation advances across symbolic and audio domains, evaluating musical quality has become a key issue and challenge. Current approaches primarily fall into three categories: subjective evaluation, which relies on human listeners but is costly and inconsistent; objective evaluation, which uses computational metrics but often fails to capture human musical perception; and combined evaluation, which integrates a mix of both processes [5].

4.1. Subjective evaluation

Subjective evaluation is generally based on the input and feedback of human listeners. Although subjective music evaluation is less reproducible (due to varying perceptions of music) and cost-friendly (caused by the need for human participants), it is still a key part of AI-generated music evaluation, as music appreciation is inherently subjective. Within current subjective evaluation approaches, the two primary methods are listener-based audio tests and visual inspection of music output.

Visual analysis of music scores often involves the involvement of an expert in music to analyse the output of symbolic music representations, including scores, chord progressions, and piano roll. For instance, one study displayed generated notes through “Rainbowgrams”, visually reconstructed as colorful spectrograms [17]. On the other hand, the music listening test is the most frequently used type of subjective evaluation. It is further split into two methods: a musical Turing Test and subjective query metrics. In subjective query metrics, participants are asked to rank aspects of computer-generated music, such as creativity, melody, and rhythm, on a numerical scale. In a musical Turing Test, AI-generated audio is compared to human-created music, with participants being asked to differentiate between the two.

4.2. Objective evaluation

Compared to the human-based subjective evaluation, objective evaluation involves the use of computational methods to assess music, producing reproducible, quantifiable indicators of quality. For example, some researchers proposed a set of simple musical metrics to evaluate and compare the output of music-generating AI models, which included the computational extraction of features including pitch, note count, average inter-onset interval (IOI), and note length [18].

4.3. Combined evaluation

When evaluating AI-generated music, we can use subjective and objective evaluation methods together to create combined evaluations. This term refers to methods that use a combination of both subjective user study and more objective, computational metrics to create a final evaluation result. For example, Huang et al. Evaluated their music mashup model, using both objective metrics and subjective listening tests [19]. By combining the two perspectives, this type of combined testing can build on the strengths of both approaches. And it helps negate each of their weaknesses, and producing a fuller and more well-rounded picture of specific models’ performances.

4.4. Challenges in evaluation

Evaluating AI-generated music remains a complex challenge, as current methods struggle with factors including standardization, consistency, and interpretability. Combined methods attempt to balance these strengths; however, aligning results across both methods remains difficult. Further complications arise in assessing creativity—novelty, originality, and value—since these qualities may be defined differently across different genres and by various audiences. Establishing unified, accurate, and efficient standards and developing metrics that combine subjective and objective evaluations are critical next steps for advancing fair and comprehensive assessment in this field [5].

5. Applications

The applications of AI music generation spans dozens of fields, showing broad cultural and commercial value. A few application areas and their brief descriptions are listed in Table 4.

Table 4. Applications for AI music generation

Content Creation	AI tools have evolved from producing simple stylistic imitations to powering large-scale creative projects. Modern systems like MuseNet and Magenta create complex, multi-genre compositions used in film, advertising, and multimedia production, which can improve both efficiency and creative flexibility in content creation and filmmaking.
Gaming and Interactive Entertainment	AI can be used to create dynamic, adaptive, and flexible soundtracks in videogames that can change in real time based on the actions of players and current in-game environments, which can help to improve immersion while also reducing soundtrack creation time and costs for game developers.
Healthcare	There has been shown to be connections between music therapy, stress relief, and rehabilitation, as well as broader mental health applications. The use of AI to create music for music therapy can create personalized, adaptive music tailored to specific individuals' needs.
Marketing	AI-generated music can strengthen brand identity by creating distinctive audio signatures and adaptive campaign soundtracks. Beyond background music, interactive ad experiences allow music to respond to consumer engagement, increasing emotional impact and brand recognition.

Across healthcare, education, media, entertainment, and the arts, AI music generation has moved from experimental novelty to widely adopted tool. By improving efficiency, enabling personalization, and inspiring further artistic innovation, it continues to shape industries and musical traditions throughout the world.

6. Challenges and future research

Despite notable advances in AI-driven music generation, especially recently, many challenges and limitations still remain, offering grounds for future research.

6.1. Challenges

Generating truly novel and varied music continues to be a major hurdle. Early systems, like David Cope's Experiments in Musical Intelligence, were able to imitate existing styles but lacked innovative and stylistically creative output. Later deep learning approaches, including GANs and RNNs, improved variety, but still often produce compositions that are overly similar in style. This is caused partially due to a phenomenon known as "model collapse", in which the training of models on their own output eventually leads to the degeneration and lack of distinctiveness in the final outputs [20].

The assessment of AI-generated music is made more difficult by inconsistencies in current evaluation methods. Traditional approaches rely on human judgment, which are often considered more reliable, but it can be subjective and lack reliability and reproducibility. On the other way, new and quantitative metrics based on statistical analysis have emerged; however, they should not be relied on by themselves and integrating them effectively with subjective human assessment is still a challenge [17].

6.2. Future research directions

In order to overcome existing hurdles in generative music technology, future research can prioritize several key directions [2]. First, more advanced types of music representations and generation techniques need to be developed and advanced. Second, interdisciplinary approaches, which involves the combining of concepts in fields such as music theory, cognitive science, and deep learning can also be leveraged, are important for advancing AI music generation, and should be explored. The development of more interdisciplinary models can improve models' ability to handle more intricate musical structures and produce outputs aligning with the aesthetic and emotional qualities traditionally desired by human listeners. Finally, more work could be done on real-time music generation and modification. It could help enhance audience engagement and participation, offering increased flexibility and creative freedom for both listeners and musicians, especially with regards to interactive music or live performances.

If steps are continuing to be taken toward the development and implementation of these areas, AI music generation is set to overcome many current limitations, achieving greater structural coherence, expressive depth, and diversity. And this progress will open numerous new doors for music creation and production, driving the rapid intelligent evolution of the use of AI in music and the fields of artificial intelligence and computer science.

7. Conclusion

This paper provides a comprehensive review of the development of AI music generation and its history, representation methods, generative models, and evaluation strategies. It follows the history of computer music from its initial origins and early applications to modern music production and the use and development of artificial intelligence with regards to it. By analyzing current models from both symbolic and audio-based approaches, as well as offering insights into sub-classifications of the two, such as diffusion models and GANs, this paper offers a structured overview of the current technological landscapes and the various types of AI models used to create music.

However, despite progress, significant challenges still remain. Current subjective, objective, and combined methods of evaluating music output by AI programs are still flawed, and quality standardized benchmarks have not yet been established. Dataset limitations, both in size and diversity, also constrain the ability of current models. Future research should therefore focus on several key directions. For one, developing more diverse and advanced datasets is essential. Additionally, advances in music representation may allow models to more faithfully reproduce the complexity of human composition, while new generative methods can help preserve originality and avoid stylistic repetition. Real-time generation and adaptive interaction could also be improved, which would enable AI systems to function as creative tools in performance and production.

In conclusion, AI music generation is a rapidly growing field with significant implications in both technology and the arts. By consolidating previously existing research and identifying current challenges that could be addressed, this paper provides a framework for future inquiry and development. As methods become more sophisticated and applications expand, AI may transform the way music is made by redefining the role of technology in human creativity and turning it from a purely human practice into a collaborative endeavor between humans and machines.

References

- [1] Pinch, T., & Trocco, F. (2004). *Analog Days: The Invention and Impact of the Moog Synthesizer*. Harvard University Press. 10.4159/9780674042162

- [2] Chen, Y., Huang, L., & Gou, T. (2024, September 3). Applications and Advances of Artificial Intelligence in Music Generation: A Review. arXiv: 2409.03715
- [3] Yang, Y. (2024, March 13). Analysis Of Different Types of Digital Audio Workstations. Highlights in Science Engineering and Technology, 85, 563-569. 10.54097/6vvy8z41
- [4] Manan, Verma, G., Singh, S., & Kukreti, K. (2022, December 23-24). A Review of Multimedia Processing and its Application in MIDI. 2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT), 1-5. 10.1109/CISCT55310.2022.10046532
- [5] Xiong, Z., Wang, W., Yu, J., Lin, Y., & Wang, Z. (2023, August 26). A Comprehensive Survey for Evaluation Methodologies of AI-Generated Music. arXiv: 2308.13736
- [6] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016, September 19). WaveNet: A Generative Model for Raw Audio. arXiv: 1609.03499
- [7] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017, April 6). Tacotron: Towards End-to-End Speech Synthesis. arXiv: 1703.10135
- [8] Schneider, F., Kamal, O., Jin, Z., & Schölkopf, B. (2024, August). Moûsai: Efficient Text-to-Music Diffusion Models (L.-W. Ku, A. Martins, & V. Srikumar, Eds.). Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). arXiv:2301.11757
- [9] Huang, Q., Park, D. S., Wang, T., Denk, T. I., Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C., Engel, J., Le, Q. V., Chan, W., Chen, Z., & Han, W. (2023, March 6). Noise2Music: Text-conditioned Music Generation with Diffusion Models. arXiv:2302.03917
- [10] Donahue, C., McAuley, J., & Puckette, M. (2019, February 9). Adversarial Audio Synthesis. arXiv:1802.04208
- [11] Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., & Yang, Y.-H. (2017, November 24). MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. arXiv:1709.06298
- [12] Liang, F. (2016, August). BachBot: Automatic composition in the style of Bach chorales. https://www.mlmi.eng.cam.ac.uk/files/feynman_liang_8224771_assignsubmission_file_liangfeynmanthesis.pdf
- [13] Hadjeres, G., Pachet, F., & Nielsen, F. (2017, June 17). DeepBach: a Steerable Model for Bach Chorales Generation. arXiv:1612.01010
- [14] Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzett, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., & Frank, C. (2023, January 26). MusicLM: Generating Music From Text. arXiv: 2301.11325
- [15] Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018, December 12). Music Transformer. arXiv: 1809.04281
- [16] Brunner, G., Konrad, A., Wang, Y., & Wattenhofer, R. (2018, September 20). MIDI-VAE: Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer. arXiv: 1809.07600
- [17] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Eck, D., Simonyan, K., & Norouzi, M. (2017, April 5). Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. arXiv: 1704.01279
- [18] Yang, L.-C., & Lerch, A. (2018, November 3). On the evaluation of generative models in music. Neural Computing and Applications, 32, 4773–4784. 10.1007/s00521-018-3849-7
- [19] Huang, J., Wang, J.-C., Smith, J. B. L., Song, X., & Wang, Y. (2021, March 26). Modeling the Compatibility of Stem Tracks to Generate Music Mashups. arXiv: 2103.14208
- [20] Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korabk, T., Agrawal, R., Pai, D., Gromov, A., Roberts, D. A., Yang, D., Donoho, D. L., & Koyejo, S. (2024, April 29). Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. arXiv: 2404.01413