

# Analysis and Comparison of Algorithmic Composition Using Transformer-Based Models

Shaozhi Pi<sup>a</sup>

*Daniel J. Epstein Department of Industrial & Systems Engineering,  
University of Southern California, Los Angeles, U.S.A.*

**Keywords:** Transformer, Music Generation, Multitrack Composition, Generative AI.

**Abstract:** As a matter of fact, transformers have revolutionized generative music by overcoming the limitations of earlier models (e.g., RNNs) in recent years, which struggled with long-term dependencies. With this in mind, this paper explores and compares four transformer-based models, i.e., Transformer-VAE, Multitrack Music Transformer, MuseGAN as well as Pop Music Transformer. To be specific, the Transformer-VAE offers hierarchical control for generating coherent long-term compositions. In addition, the Multitrack Music Transformer excels in real-time multitrack music generation with efficient memory use. At the same time, MuseGAN supports human-AI collaboration by generating multitrack music based on user input, while Pop Music Transformer focuses on rhythmic and harmonic structures, making it ideal for pop genres. According to the analysis, despite their strengths, these models face computational complexity, limited genre adaptability, and synchronization issues. Prospective advancements, including reinforcement learning and multimodal integration, are expected to enhance creative flexibility and emotional expressiveness in AI-generated music.


## 1 INTRODUCTION

Generative music, rooted in the early work of algorithmic composition, has seen considerable growth over the past few decades. Initially, techniques such as rule-based systems dominated the field (Ames, 1987; Ames, 1989). These methods were innovative but limited in capturing human-composed music's complexity, emotional depth, and creativity. However, the advent of machine learning, and more recently, deep learning, has significantly advanced the capabilities of algorithmic composition.

Early deep neural networks' approaches to generative music, such as recurrent neural networks, focused on modeling the sequential nature of music. While effective at capturing short-term dependencies, these models struggled with maintaining long-term coherence, a crucial aspect of human creativity in music. The introduction of the Transformer model by Vaswani et al., initially developed for natural language processing, revolutionized the ability of AI to generate music with complex structures (Vaswani et al, 2017). With its self-attention mechanism, this model allowed for preserving long-term

dependencies in sequential data, making it ideal for symbolic music generation. One of the most significant breakthroughs in this domain was the Music Transformer (Huang et al., 2018), which utilized self-attention to maintain coherence in music generation across extended sequences. This innovation marked a turning point in generating music that exhibits structural integrity and creativity and approaches human standards. Based on this foundation, several models have been developed to enhance specific aspects of music generation, such as multitrack music and hierarchical structures.

The Multitrack Music Transformer introduced by Dong et al., addressed the challenges of generating multitrack compositions (Dong et al, 2023). These challenges are particularly relevant for understanding the interplay of creativity across different musical elements, such as melody, harmony, and rhythm. However, while this model made significant strides in generating coherent music across multiple tracks, capturing the interdependencies between these elements remains challenging, especially in generating emotionally engaging and structurally complex music. Another significant contribution

<sup>a</sup> <https://orcid.org/0009-0001-1671-9728>

came from MuseGAN, a model that utilizes Generative Adversarial Networks (GAN) to generate multitrack music (Dong et al., 2018). Unlike transformer-based models, which focus on attention mechanisms, GAN introduce a competitive learning process that can refine the generated music through adversarial training. MuseGAN demonstrated how AI could collaboratively generate music alongside human input, with separate generators for each musical track. Repetition, structure, and emotional expressiveness are central to the human perception of creativity in music. Dai et al. noted that while deep learning models generating musically interesting sequences, they often struggle with replicating the nuanced repetition and variations found in human-composed music (Dai et al, 2022). The Transformer Variational Auto-Encoder (Transformer-VAE) was introduced to address this limitation, combining the strengths of transformer architectures with the probabilistic modeling capabilities of VAE to generate music that is both structurally coherent and emotionally expressive (Jiang et al, 2020).

As generative music continues to evolve, new models such as Jukebox by OpenAI and MusicLM by Google Research have further pushed the boundaries of AI's creative potential. Jukebox generates raw audio and is capable of vocal synthesis across various genres (Prafulla et al, 2020). In contrast, MusicLM generates music from text descriptions, combining pre-trained models to achieve stylistic and emotional diversity (Andrea et al, 2023). These advancements illustrate the rapid progress in generating music that adheres to specific creative intentions, opening new avenues for technical improvements and philosophical considerations of creativity in machine-generated music.

Despite the significant breakthroughs in music generation, challenges persist. Current models often struggle to fully capture the intricacies of musical creativity, particularly in how they integrate complex musical elements like melody, harmony, and rhythm. This underscores the need for continuous improvement and innovation in the field. Future work will likely focus on enhancing transformer-based models to capture these interactions more effectively, potentially redefining the understanding of creativity in music generation.

This paper evaluates the state-of-the-art transformer-based models for music generation, focusing on their strengths, limitations, and contributions to the broader understanding of AI-driven creativity. The paper aims to comprehensively analyze these models, including the Transformer-VAE, Multitrack Music Transformer, MuseGAN,

and Pop Music Transformer. It examines their approach to multitrack generation, structural coherence, emotional expressiveness, and technical efficiency. The goal is to offer insights into how these models have shaped the landscape of algorithmic composition and to suggest future developments that might enhance the creative capabilities of generative models.

## 2 DESCRIPTIONS OF MUSIC COMPOSING MODELS

Music generation using transformer models has gained significant attention in recent years due to the transformer's ability to model long-range dependencies within sequential data. Unlike earlier models, such as recurrent neural networks (RNN) and convolutional neural networks (CNN), which face challenges in capturing long-term dependencies and contextual relationships, transformers utilize self-attention mechanisms to process entire sequences simultaneously, making them particularly effective for music generation. The fundamental principle behind transformer models is the self-attention mechanism (Vaswani et al, 2017), which allows each element in a sequence (in this case, a musical note or event) to attend to every other element, regardless of its position in the sequence. This ability to capture dependencies across a range of time steps enables transformers to maintain coherence over extended periods, a critical feature in generating complex musical structures. The model's architecture typically consists of multiple layers of attention heads, each focusing on different aspects of the sequence, such as rhythmic patterns or harmonic progressions. This multi-layered approach ensures that the model captures local musical relationships and global structural patterns.

Symbolic music is often represented in tokenized form to apply transformers to music generation. Each token may represent a note's pitch, duration, velocity, or other musical attributes, such as instrument or articulation. This tokenization allows the transformer to process music similarly to text sequences, treating each note or event as a word in a sentence. For example, in the REMI (revamped MIDI-derived events) format, music is broken down into time events, position events, pitch events, and other performance-based tokens, which the transformer processes to generate a coherent musical output (Huang et al., 2020).

The training process for transformer models in music generation typically involves feeding large datasets of symbolic music, such as MIDI files, into the model. These datasets may represent various genres, styles, and compositional forms, allowing the transformer to learn diverse musical structures. By learning the relationships between different musical elements, the transformer can generate new compositions that reflect the micro-level details (e.g., note transitions, dynamics) and macro-level structures (e.g., harmonic progressions, form). The self-attention mechanism is particularly well-suited for capturing these layers of detail because it enables the model to focus on both close and distant relationships within the music, such as how a chord progression develops over several bars or how rhythmic patterns evolve throughout a piece.

Furthermore, transformer models use positional encoding to maintain information about the order of musical events, which is critical for generating coherent musical sequences. Since the transformer architecture does not inherently understand the sequential nature of time, positional encodings are added to each input token, enabling the model to discern the temporal structure of the music. These encodings allow the transformer to generate music that makes sense note-to-note and maintains structural integrity across longer sequences.

Another crucial aspect of music generation using transformers is autoregressive modeling. In this approach, the model generates one token at a time, predicting the next token based on the previous ones. This method ensures that the generated music remains contextually consistent, as each generated note or event is conditioned on the preceding sequence. In some cases, beam search or top-k sampling strategies ensure that the model generates more diverse and musically interesting outputs rather than simply predicting the most likely following note at each step.

Overall, transformers have demonstrated remarkable success in generating music exhibiting local coherence (e.g., smooth transitions between notes) and global structure (e.g., adherence to musical form). These models can be further enhanced by integrating additional deep learning techniques, such as VAE and GAN, which help to capture even more nuanced aspects of musical creativity, including variability and expressiveness. As a result, transformer-based models are now at the forefront of computational music generation, combining advanced deep-learning techniques with the flexibility of symbolic music representation. These models have enabled the creation of music that not only mimics traditional compositional forms but also

explores new creative possibilities that challenge conventional boundaries.

In addition to the core transformer architecture, recent advancements have introduced hierarchical models that enhance the generation of complex musical structures. For example, by utilizing local (measure-level) and global (phrase or section-level) representations, these models allow the transformer to understand the relationships between different composition parts better. This approach helps generate music that maintains thematic development and variation across extended periods, ensuring that compositions are more than just a series of disjointed musical events.

Lastly, transformer models' adaptability to various music styles and forms is another significant benefit. Whether generating classical symphonies, modern pop music, or experimental electronic compositions, transformers can be fine-tuned to learn the nuances of different genres. This flexibility is crucial for applications where creative diversity and stylistic fidelity are essential, such as music production, game soundtracks, and collaborative AI-driven composition tools.

In summary, transformer models in music composition have opened up new possibilities for generating technically proficient and creatively expressive music. By leveraging the self-attention mechanism, positional encoding, autoregressive prediction, and hierarchical structuring, these models can produce music that exhibits detailed intricacies and larger-scale coherence, making them powerful tools for advancing AI-driven music composition.

### 3 REALIZATION OF ALGORITHMS

This section will explore the technical realization of four major transformer-based models used for music generation: the Transformer-VAE, Multitrack Music Transformer, MuseGAN, and the Pop Music Transformer. The copyright form is located on the authors' reserved area.

#### 3.1 Transformer-VAE

The Transformer-VAE aims to combine the advantages of VAE and transformers to generate music that is structurally coherent but also interpretable and flexible in latent space. The core idea is to use a hierarchical model that first encodes the local structure of music (e.g., measures) and then

applies transformer layers to capture global dependencies.

In the Transformer-VAE model, music is divided into bars, and each bar is encoded into a latent representation during the Input Encoding stage. This is achieved by using a local encoder to capture the essential musical features of each bar. The Global Representation is generated by passing these bar-level latent representations through the transformer encoder. The encoder applies masked self-attention, allowing the model to understand and capture the relationships and dependencies between different bars, thus creating a coherent global structure. During the Latent Space Sampling stage, a latent code is generated for each bar based on mean and variance estimations provided by the VAE, which adds variability and creative flexibility to the music generation process. For the Music Reconstruction phase, the transformer decoder, conditioned on previously generated bars and their corresponding latent variables, reconstructs the full music sequence, ensuring continuity and coherence throughout the composition. Finally, the model allows for Context Transfer, enabling users to modify specific portions of the generated music while maintaining the overall structural integrity, offering creative control and flexibility. A flow chart is given in Fig. 1.

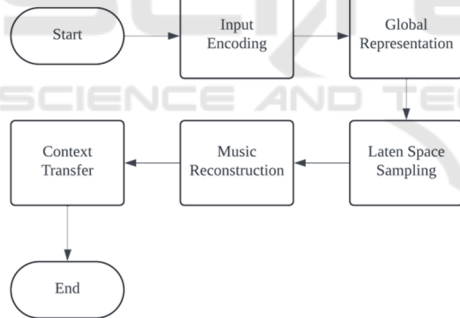


Figure 1: Algorithmic flow for Transformer-VAE (Photo/Picture credit: Original).

### 3.2 Multitrack Music Transformer

The Multitrack Music Transformer is designed to handle complex multitrack compositions while optimizing memory usage. It employs a decoder-only transformer architecture, using multidimensional input/output spaces to process each track separately. Due to its efficient architecture, this model excels in scenarios where real-time or near-real-time music generation is required.

In the Multitrack Music Transformer, each music event is represented as a tuple, which includes attributes such as note type, pitch, duration,

instrument, beat, and position. This comprehensive Data Representation allows the model to capture all relevant aspects of a musical event across multiple tracks. During the Sequence Encoding stage, these tuples are fed into the transformer decoder, where the self-attention layers process the sequences. The self-attention mechanism helps the model understand the relationships between musical events across different tracks. The transformer uses an Event Prediction mechanism for each event, following an autoregressive approach to predict the next event based on the sequence of prior events. This ensures that the generated music evolves logically over time. Lastly, Multitrack Coordination is essential for maintaining harmonic and rhythmic dependencies between instruments. The model ensures that the tracks are generated and coordinated, respecting the relationships between instruments to create a coherent multitrack composition. A flow chart is shown in Fig. 2

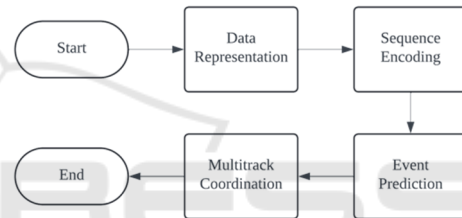


Figure 2: Algorithmic flow for Multitrack Music Transformer (Photo/Picture credit: Original).

### 3.3 MuseGAN

MuseGAN employs Generative Adversarial Networks (GAN) to generate multitrack music, focusing on separate generators for each track. This model allows for both fully automatic generation and human-AI cooperative composition, where a human provides one or more tracks, and the model generates the accompanying tracks.

In MuseGAN, each track, such as bass, drums, and piano, is generated using Track-wise Generators, where separate GAN are trained for each track. Each generator creates a piano roll for its respective track, starting from a random noise vector input. The generated tracks are evaluated by a Discriminator, which determines whether the music generated by the model is real or fake, improving the model's ability to produce authentic-sounding music. During the Training Process, the generator continuously attempts to fool the discriminator, while the discriminator becomes more adept at distinguishing between generated and real music. To maintain Multitrack



Synchronization, a shared latent vector ensures that the tracks created by different GAN are harmonized, preserving harmonic and rhythmic relationships across all tracks. Additionally, Track Conditioning allows for human-AI collaboration, where a user can input a specific track (e.g., a melody or bassline), and the model will generate the remaining tracks, conditioned on the provided input to ensure coherence and creative alignment. A flow chart is given in Fig. 3.

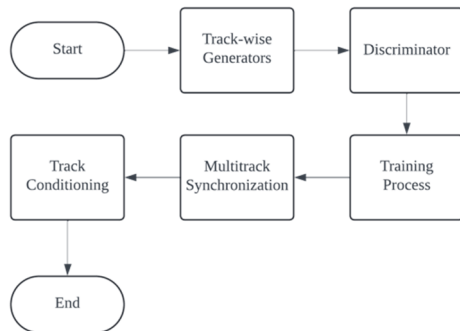


Figure 3: Algorithmic flow for MuseGAN (Photo/Picture credit: Original).

### 3.4 Pop Music Transformer

The Pop Music Transformer is explicitly designed to generate pop music that focuses on rhythm and harmonic structure. It leverages rhythmic features within the transformer's architecture, producing music that follows typical pop structures, including verses, choruses, and bridges.

In the Pop Music Transformer, the Input Representation involves tokenizing the music into rhythmic and harmonic events, emphasizing beat and meter to capture the rhythmic structure central to pop music. The model uses Positional Encoding to ensure that the temporal relationships between events are respected, allowing it to maintain the characteristic structure of pop compositions over time. During the Autoregressive Generation phase, the model generates one event at a time, predicting the next event based on the sequence of previously generated events, ensuring coherence and logical progression. To enhance the diversity and musicality of the output, Sampling Methods such as top-k or beam search are applied during generation, allowing the model to explore multiple creative pathways while staying within the bounds of the musical style. A flow chart is given in Fig. 4.

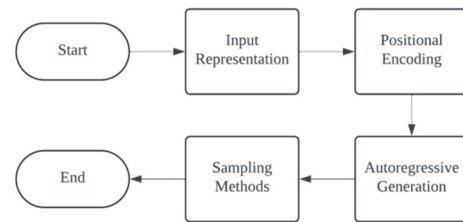


Figure 4: Algorithmic flow for Pop Music Transformer (Photo/Picture credit: Original).

## 4 COMPARISON OF MODELS

This section will compare each generation model. When comparing the four transformer-based music generation models, i.e., Transformer-VAE, Multitrack Music Transformer, MuseGAN, and Pop Music Transformer. This paper takes into account their functionality, computational efficiency, real-world use cases, and creative contributions, both in fully autonomous music generation and in collaborative human-AI processes. Each model excels in specific areas, making them more suitable for different applications depending on the user's needs.

### 4.1 Functionality

The Transformer-VAE is specialized and designed to generate long-term musical structures with flexibility and creative control. Its hierarchical model allows for both local and global structure understanding, making it ideal for compositions that must maintain thematic consistency across an extended sequence. However, its primary strength lies in providing interpretability and context transfer, allowing users to modify portions of the music and observe how it adapts.

The Multitrack Music Transformer generates complex, multitrack compositions, such as orchestral music, where multiple instruments must harmonize. Its multidimensional input/output handling allows for faster generation without sacrificing the coherence of individual tracks. Due to its efficient memory and computational use, Multitrack Music Transformer is highly suited for real-time applications like automatic accompaniment and live performances.

MuseGAN, on the other hand, is well-known for its ability to assist in both fully automated and collaborative music creation. With its track-wise generators and adversarial training, MuseGAN enables the generation of multitrack music, such as jazz bands or rock ensembles, while allowing a

human composer to control or input a specific track for accompaniment generation. This model stands out for its ability to create coherent yet distinct tracks while maintaining a harmonic balance between them.

Lastly, the Pop Music Transformer is highly optimized for generating structured, rhythmically focused compositions typical of the pop genre. It balances autonomous creativity and rhythmic control, making it an excellent tool for producers aiming to generate rhythmic loops or harmonic progressions with specific stylistic characteristics. The model's reliance on rhythmic structure ensures it stays within genre-specific rules while allowing creative flexibility.

## 4.2 Computational Efficiency

The Multitrack Music Transformer is the most optimized for computational efficiency, given its design to reduce sequence length and memory consumption. It offers faster inference times, particularly suitable for real-time or near-real-time applications. MuseGAN and Transformer-VAE, while less efficient, offer other strengths in the creative control and adaptability they afford the user. The Pop Music Transformer sits in the middle in terms of computational needs, as it balances structure and efficiency.

## 4.3 Creativity (Autonomous and Human-AI Collaboration)

Creativity is a vital aspect when comparing these models. The Transformer-VAE allows for high creativity through its hierarchical structure, enabling context transfer and offering deep insight into the relationship between different composition parts. It encourages more exploratory forms of creativity,

especially for users interested in modifying segments of generated music while maintaining overall coherence.

Multitrack Music Transformer is more focused on maintaining structure and coordination between multiple tracks, which limits its exploratory creativity but makes it highly effective in settings where harmonic and rhythmic consistency across tracks is critical. This makes it ideal for orchestrations or ensemble pieces where each part must fit together seamlessly. MuseGAN stands out for its ability to blend human and machine creativity. It is explicitly designed to assist in human-AI collaboration by allowing the user to input one or more tracks while the model generates the rest. This provides a flexible, creative process where human intuition and machine-generated content can harmonize harmoniously. While primarily focused on rhythm and harmony, the pop music transformer offers creative flexibility within the bounds of its genre. It encourages creativity in producing rhythmically structured compositions but is less adaptable to more exploratory forms of music generation.

## 4.4 Additional Metrics: User-Friendliness and Adaptability

One essential aspect not covered by computational metrics is each model's user-friendliness and adaptability. The Transformer-VAE and MuseGAN offer a higher level of user control, making them suitable for composers who want to interact directly with the generated content. In contrast, the Multitrack Music Transformer and Pop Music Transformer are more automated, offering less user input but greater efficiency in generating ready-to-use music.

Table 1: The comparison of Models.

Model	Functionality	Computational Efficiency	Creative Control	Best Use Case	Human-AI Collaboration
Transformer-VAE	Hierarchical music generation	Moderate	High: Context transfer and structure control	Long-term thematic compositions	Low
Multitrack Music Transformer	Multitrack and complex orchestration	High	Medium: Focus on track coordination	Orchestral, real-time music	Low
MuseGAN	Multitrack GAN-based generation	Low	High: User inputs one track, model generates the rest	Jazz, rock, human-AI collaboration	High
Pop Music Transformer	Rhythm and harmonic generation	Moderate	Medium: Rhythmic focus and pop structure	Pop music, Rhythmic loops	Low

## 4.5 Comparison Summary

In conclusion, the comparative analysis shows that each model has strengths and weaknesses depending on the user's specific requirements. MuseGAN is particularly useful for collaborative music generation, while Multitrack Music Transformer shines in multitrack composition for complex music. Transformer-VAE offers the most creative flexibility, especially for users interested in structural control, and Pop Music Transformer excels at generating rhythmically driven compositions with genre-specific constraints. The summaries are given in Table 1.

## 5 LIMITATIONS AND PROSPECTS

Despite the significant advancements in transformer-based models for music generation, several limitations hinder their full potential. These limitations exist both at the algorithmic level and in terms of real-world applicability, and addressing them will be critical for the future evolution of AI-generated music. One of the Transformer-VAE model's primary limitations is its complexity and computational demands. While the hierarchical structure allows for greater control and flexibility in generating long-term thematic compositions, it can be computationally expensive. Balancing local and global structures through multiple layers of encoders and decoders increases training time and requires considerable memory. Additionally, the model's reliance on latent space sampling may lead to a loss of fine-grained detail in musical generation, as it simplifies complex musical elements into latent variables that can sometimes lose nuance.

While the Multitrack Music Transformer's efficiency in handling multitrack compositions is a strength, it struggles with interdependencies between tracks when complexity increases, such as in orchestral compositions with numerous instruments. Moreover, its focus on memory optimization and faster inference times can come at the cost of creative flexibility. It excels at keeping track of coherence but is limited in exploring more innovative, experimental music.

MuseGAN, although strong in collaborative human-AI interaction, faces challenges in model evaluation. The adversarial training inherent in GAN often suffers from issues such as mode collapse, where the generator produces repetitive outputs with limited diversity. Furthermore, because the model

relies on separate generators for each track, it can sometimes fail to fully synchronize across tracks, resulting in minor harmonic or rhythmic dissonances.

The Pop Music Transformer, explicitly designed for generating rhythmic and harmonic pop music, can limit genre diversity. It excels in structured, predictable genres like pop but lacks the adaptability required for more complex or experimental compositions. Additionally, while it generates coherent music, its reliance on predefined rhythmic structures may restrict creative freedom, making it less useful for users seeking to push the boundaries of conventional music.

Looking ahead, several key areas for improvement can enhance these models' capabilities. One potential direction is integrating reinforcement learning to encourage greater musical diversity and creativity. By allowing models to explore different musical paths and receive feedback based on aesthetic or stylistic goals, it would be possible to generate more innovative and less predictable compositions. For example, models like Transformer-VAE could benefit from reinforcement learning to fine-tune the generation process, balancing the trade-off between structural coherence and creative exploration. Another promising area for improvement is the development of multimodal models that integrate audio, text, and visual inputs. By training models on datasets that combine musical scores, text descriptions, and even visual cues (e.g., music videos), models could generate music that aligns with specific artistic intentions or narratives. This would open up new possibilities for music generation, particularly in fields like film scoring and video game soundtracks, where the music must dynamically respond to visual content. Improved evaluation metrics are also essential for future research. Current models are often evaluated based on subjective listening tests or objective metrics like coherence. However, there is a need for more sophisticated metrics that can measure emotional expressiveness, creativity, and originality in AI-generated music. By developing better tools for assessing these qualities, researchers can improve models like MuseGAN and Pop Music Transformer, allowing them to generate music that follows structural rules and evokes a deeper emotional response.

In conclusion, while transformer-based models for music generation have made great strides, they could be better. Addressing the limitations related to computational efficiency, inter-track dependencies, genre diversity, and creative freedom will be crucial to advancing the field. Future developments are likely to focus on integrating reinforcement learning,

expanding to multimodal inputs, and refining evaluation metrics to create music that is not only technically proficient but also emotionally and creatively compelling.

## 6 CONCLUSION

To sum up, this paper has explored the functionality, computational efficiency, and creative potential of four key transformer-based models for music generation: Transformer-VAE, Multitrack Music Transformer, MuseGAN, and Pop Music Transformer. Each model has its strengths and limitations, from the flexibility and structural control of Transformer-VAE to the efficiency and multitrack harmony capabilities of the Multitrack Music Transformer. MuseGAN excels in human-AI collaboration, while the Pop Music Transformer generates rhythmically focused compositions suitable for pop music. These models showcase how AI can contribute to the creative process by generating coherent and structured music across various genres and use cases. Future works in this field have great potential to expand through improvements in reinforcement learning, multimodal integration, and evaluation metrics that better capture creativity and emotional expressiveness. These advancements will further enhance the ability of AI models to produce innovative and emotionally compelling compositions. In conclusion, these results contribute to the growing body of research on algorithmic composition, highlighting the strengths and challenges of current models while identifying areas for future development. By pushing the boundaries of AI-generated music, these models represent significant strides toward bridging the gap between human creativity and machine learning.

## REFERENCES

- Agostinelli, A., Denk, T. I., Borsos, Z., et al, 2023. *Musiclm: Generating music from text*. arxiv preprint arxiv:2301.11325.
- Ames, C., 1987. *Automated Composition in Retrospect: 1956–1986*. Leonardo 20(2), 169-185. 5.
- Ames, C., 1989. *The Markov Process as a compositional Model: a survey and tutorial*. Leonardo, 22(2), 175.
- Dai, S., Yu, H., Dannenberg, R. B., 2022. *What is missing in deep music generation? A study of repetition and structure in popular music*. 23rd International Society for Music Information Retrieval Conference, 11
- Dong, H. W., Hsiao, W. Y., Yang, L. C., Yang, Y. H., 2018. *MuseGAN: multitrack sequential generative adversarial networks for symbolic music generation and accompaniment*. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium, 32, 1.
- Dong, H. W., Chen, K., Dubnov, S., McAuley, J., Berg-Kirkpatrick, T., 2023. *Multitrack music transformer*. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 1-5.
- Huang, C. Z., Vaswani, A., Uszkoreit, J., Shazeer, N., Hawthorne, C., Dai, A., Hoffman, M., Eck, D., 2018. *Music Transformer: Generating Music with Long-Term Structure*. arXiv preprint arXiv:1809.04281.
- Huang, Y. S., Yang, Y. H., 2020. *Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions*. Proceedings of the 28th ACM International Conference on Multimedia, 1180–1188.
- Jiang, J., Xia, G. G., Carlton, D. B., Anderson, C. N., Miyakawa, R. H., 2020. *Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning*. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 516-520.
- Prafulla, D., Heewoo, J., Christine, P., Jong, W. K., Alec, R., Ilya, S., 2020. *Jukebox: A Generative Model for Music*. arxiv preprint arxiv:2005.00341.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, U., Polosukhin, I., 2017. *Attention is All you Need*. Advances in Neural Information Processing Systems. Curran Associates, Inc..