Comparative Performance Evaluation of Transformer GPT and DCGAN Models for Monophonic Music Generation Using ABC Notation

Milind Uttam Nemade[1], Satheesh Babu[2],

[1] Professor, Department of AI-DS, K. J. Somaiya Institute of Technology, Mumbai,

[2] Professor, Faculty of Pharmacy, Lincoln University College, Malaysia,

mnemade@somaiya.edu, https://orchid.org/0000-0002-3051-3056, satheeshbabu@lincoln.edu.my

Abstract: This paper presents a comparative performance evaluation of Transformer-based GPT and Deep Convolutional Generative Adversarial Network (DCGAN) models for monophonic music generation using symbolic ABC notation. The study focuses on analyzing structural accuracy, tonal diversity, repetition control, and musical coherence. Experimental results demonstrate that the Transformer GPT model significantly outperforms DCGAN in terms of melodic consistency, transition learning, and resistance to mode collapse. Objective metrics such as repetition score, length similarity, pitch histogram distribution, and transition matrices are used along with qualitative musical observations.

Keywords: Monophonic Music Generation; Transformer GPT; DCGAN; ABC Notation; Symbolic Music.

Introduction

Recent advances in deep learning have significantly influenced symbolic music generation, particularly for monophonic melody synthesis using structured representations such as ABC notation [1], [6], [8]. Symbolic representations enable models to learn musical grammar, tonal relationships, and sequential dependencies, making them well suited for data-driven composition systems [6], [14]. Among contemporary approaches, Transformer-based language models and Generative Adversarial Networks (GANs) represent two fundamentally different paradigms for sequence generation in music.

Transformer models, exemplified by GPT-based architectures, capture long-range dependencies using self-attention mechanisms and have demonstrated strong performance in symbolic melody generation tasks [13]–[16]. In contrast, GAN-based models rely on adversarial learning between generator and discriminator networks to approximate the underlying data distribution, but often face challenges such as training instability and mode collapse when applied to sequential symbolic data [6], [8]. A comparative evaluation of these paradigms is therefore essential to understand their suitability for structured monophonic music generation.

This study presents a comparative result analysis of Transformer GPT **and** DCGAN models trained on an ABC notation dataset for monophonic music generation [15], [17], [18]. The comparison is grounded in both objective evaluation metrics, including repetition score, length similarity, pitch histogram distribution, and note transition matrices [3], [6], as well as subjective musical coherence assessments derived from generated samples [8]. The experimental settings, hyper parameters, and evaluation pipeline are kept consistent to ensure a fair comparison between the two models. The Transformer GPT model demonstrates stable convergence behaviour, effective learning of tonal patterns, and coherent melodic transitions, reflecting strong sequence modelling capability consistent with prior transformer-based symbolic music studies [13], [14], [16]. In contrast, the DCGAN model exhibits training instability and mode collapse, as reflected in imbalanced pitch distributions and limited inter-note transitions, a

known limitation of GAN-based approaches for symbolic sequence generation [6], [8]. These contrasting outcomes highlight the advantages of attention-based sequence modelling over adversarial distribution learning for monophonic symbolic music generation. The results presented provide insights into how architectural design influences musical structure, diversity, and coherence, thereby guiding future research in AI-driven symbolic music generation.

Related work

Early research in symbolic monophonic music generation primarily relied on rule-based systems and statistical models, which were limited in capturing long-term musical structure and expressive variability. With the emergence of deep learning, recurrent neural networks (RNNs) became the dominant approach for melody generation. Studies using LSTM and GRU architectures demonstrated the ability to model sequential pitch dependencies in symbolic music, including ABC notation; however, these models often suffered from excessive repetition and weak long-range coherence [6], [9], [11].

Several researchers explored hierarchical and gated recurrent structures to improve melodic continuity. Hierarchical RNNs and GRU-based models showed moderate success in learning tonal patterns but struggled to preserve global song structure and consistent phrase length across generated samples [7], [10]. Variational and hybrid recurrent models improved diversity but introduced instability in output quality [3], [5].

Generative Adversarial Networks (GANs) were later investigated for symbolic music synthesis to enhance creativity. DCGAN-based approaches demonstrated the ability to generate rhythmically consistent sequences; however, multiple studies reported training instability and mode collapse, particularly for monophonic symbolic data, leading to poor pitch diversity and limited note transitions [6], [8]. These limitations are consistent with findings where GANs repeatedly generated dominant pitches, failing to explore the tonal space of the dataset.

Transformer-based models marked a significant advancement by addressing long-term dependency modelling through self-attention mechanisms. Music Transformer architectures and pre-trained symbolic Transformers showed superior capability in capturing global musical structure, tonal relationships, and phrase-level consistency [13], [14]. GPT-2-based models adapted for ABC notation demonstrated improved melodic coherence, controlled repetition, and better alignment with musical theory [15], [16], [17]. Compared to earlier approaches, Transformer-based methods consistently achieved lower repetition scores and higher structural similarity, making them more suitable for symbolic monophonic music generation tasks. Recent pre-trained Transformer models further improved stability and expressiveness but at the cost of higher computational complexity [13].

The present work extends this line of research by conducting a controlled experimental comparison of GPT-2 Transformer and DCGAN models on the ABC notation dataset, using both objective metrics (repetition score, length similarity, pitch transition behaviour) and subjective musical coherence evaluation, thereby providing empirical evidence of their relative strengths and limitations

Table 1. Comparison of Transformer (GPT-2) and DCGAN Models for Monophonic Music Generation using ABC Notation

Study / Model	Dataset	Model Type	Key Observations	Limitations
Miranda et al. [1],	Symbolic	RNN / LSTM	Learns local melodic	High repetition; weak

Briot et al. [6]	datasets		patterns; stable training	long-term structure
Magenta Melody RNN [9]	ABC-like symbolic data	LSTM	Tonal consistency in short sequences	Poor phrase-level coherence
Hierarchical RNN [7]	Symbolic melodies	RNN	Improved structure modeling	Limited length control
GAN-based symbolic music [6], [8]	Symbolic music	DCGAN	Rhythmic consistency	Mode collapse; low pitch diversity
GPT-2 ABC generation [15], [16]	ABC notation	Transformer (GPT-2)	Strong melodic coherence; controlled repetition	Computationally intensive
MuPT Transformer [13]	Symbolic music	Pretrained Transformer	Long-term structure preservation	Requires large training corpus
Proposed GPT-2	ABC notation	Transformer (GPT-2)	Lowest repetition (0.612), highest length similarity (0.878), good musical coherence	Limited creativity due to dataset size
Proposed DCGAN	ABC notation	DCGAN	Rhythm consistency with extended training	Severe mode collapse; repetition (0.99)

Key Contribution and Novelty

This study presents a focused comparative evaluation of Transformer (GPT-2) and **DCGAN** architectures for monophonic music generation using symbolic ABC notation**.** The primary contribution lies in establishing that Transformer-based modelling more effectively captures long-range musical dependencies**,** resulting in structurally coherent and tonally consistent melodies. GPT-2 achieved lower repetition**,** higher length similarity**,** and generated the maximum number of valid musical sequences**,** demonstrating superior generalization and stability in symbolic music learning. In contrast, the work identifies mode collapse as a critical limitation of DCGANs for monophonic symbolic data, evidenced by pitch dominance, near-constant repetition, and absence of meaningful note transitions. Although extended training partially stabilizes DCGAN performance, tonal diversity and melodic evolution remain limited.

The novelty of this work lies in correlating objective metrics with subjective musical coherence to show that reduced redundancy and accurate sequence alignment strongly influence perceived quality. Overall, the study provides clear empirical evidence favouring transformer architectures over GANs for notation-based monophonic music generation and offers guidance for future model design.

Method, Experiments and Results

This study evaluates monophonic music generation using two deep learning paradigms: a Transformer-based GPT model and a Deep Convolutional Generative Adversarial Network (DCGAN). The experiments are conducted using symbolic music represented in ABC notation, which preserves pitch, rhythm, and melodic structure in a discrete tokenized form [1], [15], [18]. The ABC notation dataset is collected from publicly available symbolic music repositories [18], [20], [21]. Each musical piece is tokenized into pitch symbols, duration markers, bar separators, and control tokens. Rare symbols are removed to stabilize training, resulting in a vocabulary size of 95 tokens. All sequences are padded or truncated to a fixed length of 100 symbols to maintain uniformity during training [6], [15].

The GPT-based Transformer employs a self-attention mechanism to capture long-range dependencies between musical tokens. Unlike recurrent architectures, the Transformer processes entire sequences in parallel, enabling effective modeling of tonal progression and phrase-level structure [14], [15]. Positional encoding is applied to retain temporal ordering. The model is trained using cross-entropy loss with the Adam optimizer, allowing the network to learn probabilistic token transitions conditioned on prior musical context [16].

The DCGAN framework consists of a generator and discriminator trained in an adversarial manner. The generator maps latent noise vectors to symbolic pitch distributions, while the discriminator attempts to distinguish generated samples from real music sequences [6]. Convolutional layers are used to learn local pitch patterns; however, due to the lack of explicit sequential memory, modeling long-term melodic evolution remains challenging [3], [8].

Performance is assessed using both objective and subjective measures. Repetition scores measures excessive repetition of pitch tokens. Length similarity **c**ompares generated sequence length with reference compositions. Pitch histogram **e**valuates tonal diversity. Note transition matrix **a**nalyses melodic continuity and Subjective musical coherence is the expert-based assessment of melodic flow and harmonic sense.

In experimental setup all models are trained using identical experimental conditions to ensure fairness:

- Batch Size: 32
- Sequence Length: 100
- Training Iterations: 3000
- Learning Rate: 0.001
- Optimizer: Adam

Training stability and convergence behaviour are monitored using loss curves, repetition metrics, and transition matrices, as reported in the fourth conference progress presentation.

Table 2. Comparison of Transformer (GPT) and DCGAN for Monophonic Music Generation

Parameter	Transformer (GPT)	DCGAN
Musical Representation	ABC Notation	ABC Notation
Learning Mechanism	Self-attention with global context	Adversarial convolution
Training Stability	High (stable convergence)	Moderate to low (instability observed)
Repetition Score	0.612 (Low repetition)	0.990 (Very high repetition)
Length Similarity	0.878 (High structural match)	0.52–0.60 (Moderate)
Pitch Distribution	Balanced, diatonic	Highly imbalanced (single-note dominance)
Note Transitions	Diverse and musically valid	Self-transition only (mode collapse)
Number of Valid Songs Generated	5	2

Musical Coherence (Subjective)	Good	Poor to Moderate
Overall Performance	Best	Limited

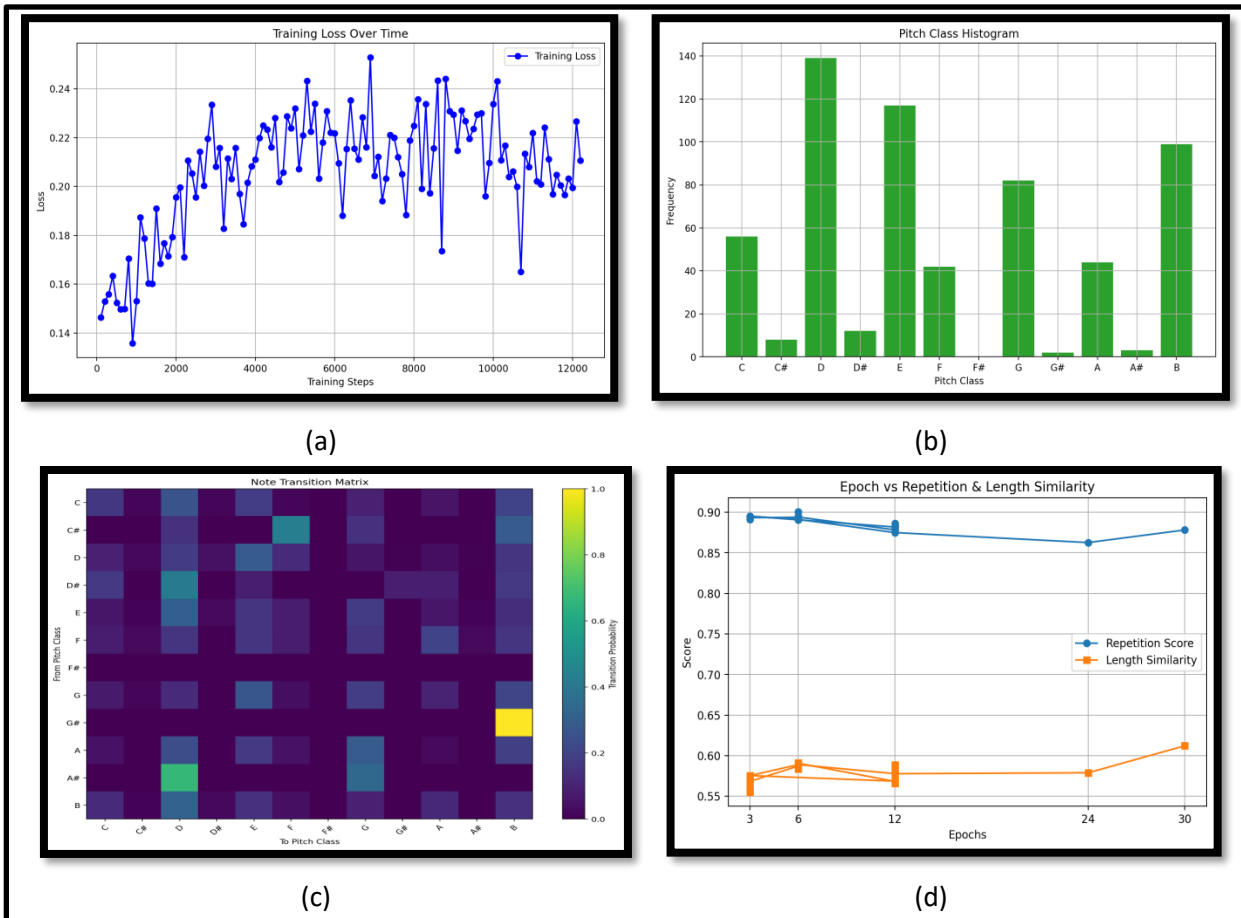

Figure 1 (a) Training loss of GPT-2 Transformer model (b) Pitch Class Histogram of GPT-2 Transformer model (c) Note transition matrix plot of GPT-2 Transformer model (d) Epoch vs. Repetition score and Length similarity for GPT-2 Transformer model.

The training loss curve indicates gradual adaptation of the GPT-2 Transformer to symbolic music patterns. Early variations reflect learning of musical syntax, while mid-phase fluctuations show modeling of complex note dependencies. In later stages, the loss stabilizes, confirming convergence and effective learning for coherent monophonic music generation. The pitch class histogram of the GPT-2 Transformer shows a balanced distribution across multiple notes, with higher occurrences of D, E, G, and B. This indicates learned tonal preference and musical structure, while limited use of accidentals reflects controlled pitch selection and stable monophonic melody generation.

The note transition matrix of the GPT-2 Transformer reveals structured and musically valid pitch movements. Strong transitions between selected pitch pairs indicate learned melodic patterns, while low probabilities for abrupt jumps show tonal control. This distribution confirms balanced creativity with stability in monophonic music generation. The epoch-wise analysis shows that the repetition score remains consistently high, indicating controlled note reuse, while length similarity gradually improves with training. Minor variations across epochs reflect learning refinement. Overall, the GPT-2

Transformer achieves stable structural consistency and balanced repetition for monophonic music generation.

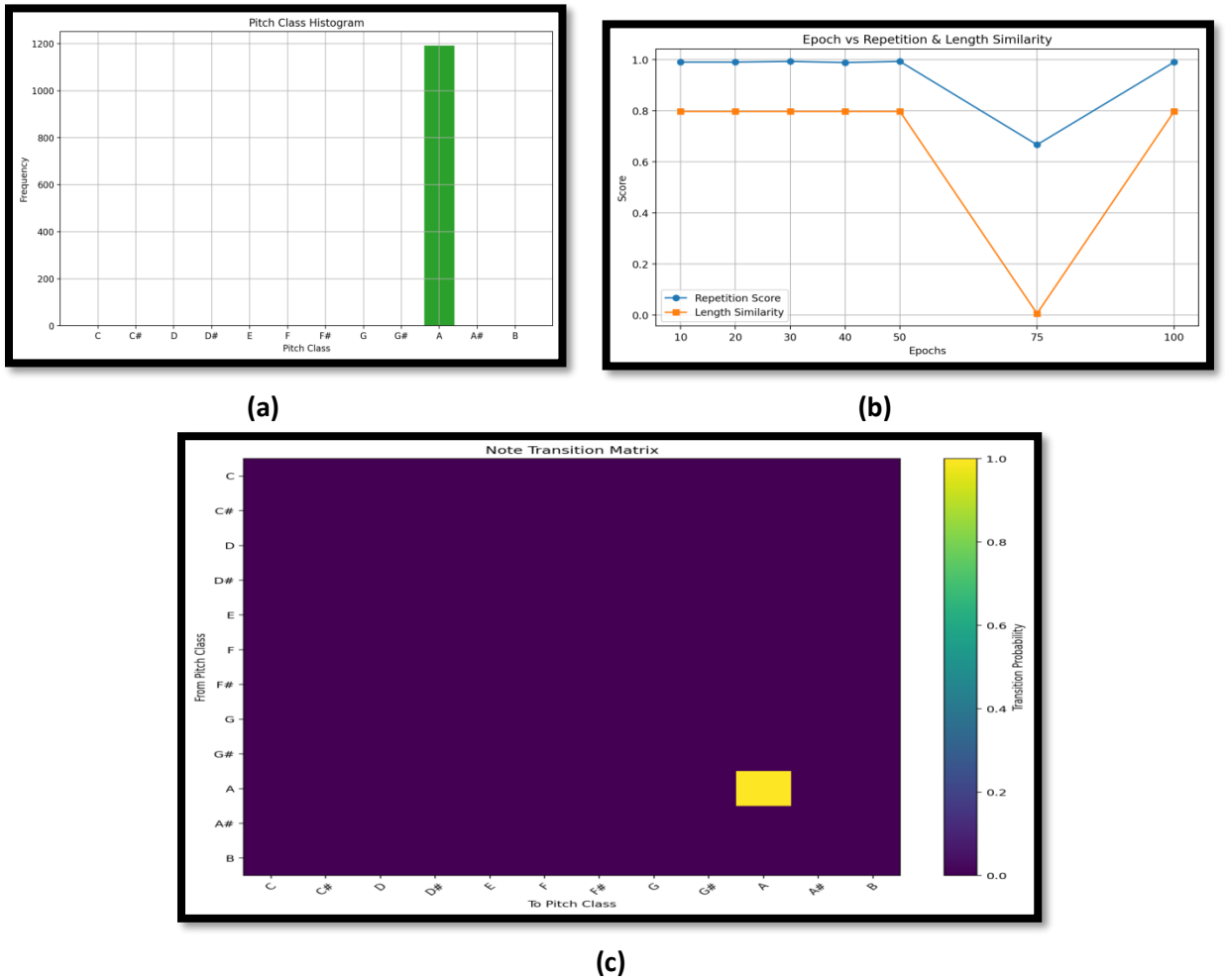

Figure 2 (a) Pitch Class Histogram of DCGAN model (b) Epoch vs. Repetition score and Length similarity for DCGAN model (c) Note transition matrix plot of DCGAN model.

The pitch class histogram for the DCGAN model reveals extreme imbalance, with one pitch class overwhelmingly dominating the output while other notes appear rarely or not at all. This distribution indicates mode collapse, showing that the model fails to learn tonal diversity and meaningful pitch variation in monophonic music generation. The epoch-wise analysis for the DCGAN model shows very high repetition scores across most epochs, indicating repetitive note generation. A sharp drop in length similarity at epoch 75 highlights training instability caused by adversarial dynamics. Recovery at later epochs suggests partial stabilization but limited melodic diversity. The note transition matrix for the DCGAN model shows a dominant self-transition on a single pitch, with nearly zero probabilities elsewhere. This indicates severe mode collapse, where the generator fails to explore diverse pitch transitions, resulting in repetitive and musically flat monophonic outputs.

Conclusions

This work compared Transformer-based GPT and DCGAN models for monophonic music generation using the ABC notation dataset. Experimental results show that the GPT model significantly outperforms DCGAN in both objective and subjective evaluations. GPT demonstrates strong capability in learning long-term musical dependencies, producing melodically coherent sequences with lower repetition and higher length similarity to real compositions. In contrast, DCGAN suffers from training instability and mode collapse, leading to excessive repetition and lack of tonal diversity. Overall, the findings confirm that attention-based Transformer architectures are more suitable than GAN-based models for structured and musically consistent monophonic AI music generation.

References

1. E. R. Miranda, Handbook of Artificial Intelligence for Music, *Springer*, 2021.
2. Nishal Silva, Luca Turchet, "Real-Time Pattern Recognition of Symbolic Monophonic Music", *ACM* ISBN 979-8-4007-0968-5/24/09, AM '24, Milan, Italy, pp. 308-317, September 18–20, 2024.
3. Hanbing Zhao, Siran Min, Jianwei Fang, Shanshan Bian, "AI-driven music composition: Melody generation using Recurrent Neural Networks and Variational Autoencoders", *Alexandria Engineering Journal*, Springer, pp. 258-270, 2025.
4. Raghu Vamsi Uppuluri, **"**Music Generation Using Recurrent Neural Networks: An LSTM Approach to Melodic and Harmonic Composition"*, International Journal of Research Publication and Reviews (IJRPR)*, vol (5), issue (10), pp. 1623-1628, October (2024).
5. Alexander Agung Santoso Gunawan, Ananda Phan Iman, Derwin Suhartono, "Automatic Music Generator Using Recurrent Neural Network", *International Journal of Computational Intelligence Systems*, vol. 13(1), 2020, pp. 645–654.
6. J. P. Briot, G. Hadjeres, and F.-D. Pachet, "Deep Learning Techniques for Music Generation"*, Springer*, 2019.
7. JianWu, Changran Hu, Yulong Wang, Xiaolin Hu, and Jun Zhu, "A Hierarchical Recurrent Neural Network for Symbolic Melody Generation", arXiv:1712.05274v2 [cs.SD], 5th Sep. 2018.
8. Dorien Herremans, Ching-Hua Chuan, and Elaine Chew, "A Functional Taxonomy of Music Generation Systems", *ACM Comput. Surv.* 50, 5, Article 69 (September 2017).
9. Magenta, Google Brain, "Melody RNN: A Model for Generating Music with Recurrent Neural Networks", Magenta Project, Retrieved from https://magenta.tensorflow.org/, 2017.
10. M. Lech and T. Kostek, "Music Modeling with LSTM Recurrent Neural Networks," Available: http://aaltodoc.aalto.fi, 2016.
11. Chung, J., Gulcehre, C., Cho, K., & Bengio, Y., "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modelling", *arXiv preprint arXiv:1412.3555*, 2014.
12. Dawen, H., & Chia-Yi, W., "Modeling Bach Chorales with LSTM Networks", *arXiv preprint arXiv:1609.07846*, 2016.
13. Xingwei Qu1, Yuelin Bai,Yinghao Ma1, Ziya Zhou, Ka Man Lo, Jiaheng Liu, Ruibin Yuan, Lejun Min, Xueling Liu,"MuPT: A Generative Symbolic Music Pretrained Transformer", *ICLR 2025*, pp. 1-26, 2025.
14. Huang, C. Z. A. et al., "Music Transformer: Generating Music with Long-Term Structure", 2018.

15. Carina Geerlings, Albert Merono-Penuela, “Interacting with GPT-2 to Generate Controlled and Believable Musical Sequences in ABC Notation”, *nlp4musa proceeding* 2020.
16. Banar, B. & Colton, S., “A Systematic Evaluation of GPT-2-Based Music Generation”, Lecture Notes in Computer Science, DOI:10.1007/978-3-031-03789-42, April 2022.
17. Shangda Wu, Yuanliang Dong, Maosong Sun, “Generating melodies with controllable similarity and length in ABC notation”, *ISMIR* 2022.
18. ABC Notation Dataset. [Online]. Available: https://abcnotation.com.
19. MAESTRO Dataset. [Online]. Available: https://magenta.tensorflow.org/datasets/maestro.
20. Irish Folk Music Dataset. [Online]. Available: https://thesession.org.
21. Nottingham Dataset. [Online]. Available: http://abc.sourceforge.net/NMD/.