

Graph Neural Network-Based Musical Structure Modeling and Singing–Melody Relationship Analysis

Buraj Rhuja

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.

surajahuja889@missouri.edu

Ashish L. Parekh

Department of Computer Science, Binghamton University, Binghamton, NY, USA.

parekhashish@binghamton.edu

Ruben Webb

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.

webb1986@buffalo.edu

Abstract

The computational analysis of musical structure and the relationship between singing performance and notated melody has historically relied on sequential or grid-like representations that do not fully capture the relational dependencies among musical events. Graph neural networks offer a compelling alternative by modeling notes, beats, measures, segments, and vocal contours as interacting nodes in a graph, with edges encoding temporal continuity, harmonic similarity, repetition, and score-to-performance alignment. This paper presents an interdisciplinary examination of graph neural network-based musical structure modeling and singing-melody relationship analysis from a systems perspective. It discusses how graph representations can integrate audio features, symbolic score information, and singing voice characteristics within a unified relational framework. The paper addresses architectural trade-offs involving message passing schemes, node granularity, heterogeneous edge types, and scalability across large and culturally diverse corpora. Beyond algorithmic considerations, the work analyzes infrastructure deployment challenges, data governance, robustness under noisy field recordings, fairness across musical traditions, and sustainability of large-scale model training. The discussion highlights the importance of treating these systems not as isolated classifiers but as socio-technical infrastructures embedded in creative, archival, and educational contexts. The paper concludes with forward-looking perspectives on transparent evaluation, community-centered dataset governance, and energy-aware deployment for music structure and singing analysis systems.

Keywords

graph neural networks; music information retrieval; music structure analysis; singing voice processing; melody alignment; sociotechnical systems.

1. Introduction

Graph neural networks have become central to relational learning in complex domains. Early graph convolutional formulations [1] and attention-based message passing [2] established that node representations benefit from aggregating information over irregular neighborhoods rather than assuming Euclidean locality. A comprehensive survey [3] documented the

expansion of these methods across node classification, link prediction, and graph-level reasoning. In music information retrieval, such relational learning is particularly relevant because musical events derive meaning from hierarchical context, temporal recurrence, and interaction with other events. A note is not merely an isolated pitch value; it acquires function through its position within a phrase, its relation to harmony, and its participation in repeated structures. These relational properties make graph-based modeling a natural fit for music analysis tasks, especially those that connect performance data with symbolic scores.

Music structure analysis has traditionally segmented audio into verses, choruses, and bridges using self-similarity matrices and clustering. Methods based on spectral clustering [4] and time series structure features [5] produce useful segmentations but often treat structure as a flat similarity pattern. The availability of a global multi-technique singing corpus with realistic music scores [6] provides new opportunities for linking structural segmentation to singing voice characteristics and melodic notation. This link is central to the present work because singing-melody relationships are not merely frame-level pitch tracking problems. They involve systematic correspondences between notated musical events, expressive vocal gestures, linguistic syllables, and structural context. A systems perspective is thus required, one that spans representation, architecture, deployment, governance, and policy.

The remainder of this paper develops that systems perspective. Section 2 reviews relevant technical foundations. Section 3 analyzes graph neural network-based musical structure modeling. Section 4 examines singing-melody relationship analysis through relational representations. Section 5 discusses system architecture and infrastructure considerations. Section 6 addresses robustness, fairness, and governance. Section 7 explores deployment, sustainability, and policy implications. Section 8 concludes.

2. Background and Related Work

Sequence-oriented models such as the Music Transformer [7] exploited self-attention to generate coherent long-term structure but still operate over a predefined positional order. Singing melody research has developed specialized pitch estimators such as CREPE [8] and deep salience representations [9] that estimate fundamental frequency in polyphonic mixtures. Evaluation frameworks like *mir_eval* [10] and foundational treatments of music processing [11] provide shared metrics for segmentation, transcription, and alignment. Content-based music information retrieval has been framed as a multidisciplinary systems problem [12], in which representational choices affect scalability, reproducibility, and usefulness. These prior contributions provide essential components for graph-based approaches, but they do not fully address the relational integration of structure and singing.

Graph neural networks extend these foundations by allowing musical objects to be modeled as nodes connected through multiple semantically meaningful edges. Unlike a sequence, a graph can simultaneously encode temporal adjacency, metric hierarchy, harmonic similarity, and repetition. This expressive capacity is valuable for music because different structural dimensions operate at different timescales. Harmonic change may occur every measure, sectional repetition may occur every half minute, and vocal ornamentation may occur within a single syllable. A graph representation can preserve all of these relations without forcing them into a single linear order.

3. Graph Neural Network-Based Musical Structure Modeling

Graph-based models represent a musical piece as a collection of nodes associated with audio frames, beats, measures, chord regions, or notated phrases. Edges connect temporally adjacent

units, repeated segments, metrically related positions, and harmonically similar regions. Unlike self-similarity matrices, graphs make relational context explicit and enable learned propagation across different musical timescales. Systematic explorations of computational music structure research [13] have highlighted the need for mid-level representations, and graph neural networks for symbolic scores [14] have demonstrated that relational structures can improve music understanding tasks. Early audio-based structure analysis [15] established segmentation quality measures and emphasized repetition, while later learned audio representations [16] provide robust inputs for graph construction.

One major system-level trade-off is node granularity. Fine-grained nodes such as constant-Q transform frames preserve detail but produce dense graphs, while measure-level or segment-level nodes reduce computational load but may obscure expressive timing. Heterogeneous graphs distinguish between note, onset, syllable, and section nodes; however, heterogeneous message passing introduces additional parameterization and design complexity. Graph attention mechanisms can weigh edges adaptively, but they also increase memory requirements. Shared-weight graph convolutions scale more readily across long recordings, yet they may be less sensitive to rare musical events. These trade-offs must be evaluated not only by segmentation accuracy but also by training stability, inference cost, and interpretability.

Graph construction also shapes governance and fairness. If edges are based primarily on Western tonal similarity, the resulting model may impose hierarchical assumptions that do not suit monophonic, drone-based, or rhythmic traditions. A systems-oriented design therefore treats graph schema as a policy-relevant artifact. The choice of node identities and edge semantics determines which musical relations are recognized, which are ignored, and which communities are served. Transparent documentation of graph construction can support independent audits and improve reproducibility. Furthermore, graph sparsification and edge selection strategies must be designed with attention to downstream use, because removing edges can erase structural information that is critical for underrepresented musical practices.

4. Singing–Melody Relationship Analysis

Singing voice processing often requires separating the vocal line from instrumental accompaniment before melody analysis. Source separation models such as conditioned U-Net architectures [17] can use pitch or score information as control signals, thereby creating a pathway from symbolic expectations to audio streams. Within a graph framework, the separated vocal representation can be aligned to note nodes derived from a score or lead sheet. Edges then connect sung syllable nodes to corresponding notated pitch nodes, allowing the model to learn expressive deviations such as vibrato, portamento, timing variation, and ornamentation. This relational formulation is more expressive than treating singing as a continuous contour without reference to notated structure.

The relationship between singing and melody is not a simple one-to-one mapping. Singers intentionally alter pitch, onset, offset, and dynamics for stylistic and affective purposes. A graph neural network can treat score-to-performance alignment as node matching over a bipartite graph, with temporal edges enforcing order and similarity edges linking perceptually related events. This relational formulation provides stronger inductive bias than frame-level regression or sequence-to-sequence models because it explicitly represents the structural correspondence between notated and performed events. It also supports partial matching under missing or noisy data, which is common in historical recordings and informal performances.

Systems for singing melody analysis must be robust to reverb, overlapping instruments, recording quality, and informal performance settings. Graph-based models can incorporate uncertainty by allowing multiple candidate edges or learned edge weights. However, this flexibility introduces vulnerability to noisy graph construction and adversarial perturbations. Deployment in archival or educational contexts therefore requires confidence calibration, anomaly detection, and fallback mechanisms that avoid propagating silent failures. It also requires evaluation protocols that report performance separately for clean studio settings, noisy field recordings, and diverse vocal styles.

5. System Architecture and Infrastructure Considerations

Deploying graph neural network-based music structure and singing analysis tools requires attention to data pipelines, feature extraction services, graph construction modules, model serving, and downstream interfaces. Foundation model discussions [18] have emphasized that large-scale model development creates emergent infrastructure demands distinct from single-model experimentation. A music analysis system must handle heterogeneous inputs such as studio recordings, field recordings, symbolic scores, and crowd-sourced annotations. Batch processing for archival collections may prioritize throughput, while interactive educational tools require low-latency inference and incremental updates. These operational differences influence architecture choices such as caching, edge partitioning, and model replication.

Architecture should separate feature extraction, graph construction, graph learning, and post-hoc explanation. This modularity supports independent audits and replacement of components. However, modular boundaries can also introduce integration overhead. For example, upstream source separation errors propagate into graph node features and edge predictions; downstream interpretation layers must account for this error propagation. System designers should maintain provenance metadata that records feature versions, model versions, graph schema versions, and evaluation datasets. Such provenance information is essential for reproducing research results and for tracing how data governance decisions affect model behavior.

Scalability is another structural concern. Long musical works, large corpora, and multi-track recordings can produce graphs with millions of nodes and edges. Distributed graph processing or sampling strategies are necessary for training, but sampling may bias learning toward densely connected regions. System architectures must therefore include monitoring for performance degradation across duration, genre, and recording condition. Operational dashboards can track segmentation stability, alignment error, and inference latency. This monitoring supports not only engineering reliability but also accountability to the communities whose music is being processed.

6. Robustness, Fairness, and Governance

Large model training can amplify biases present in datasets [19]. In music analysis, bias may manifest as poor performance on non-Western scales, underrepresented languages, and vocal styles outside mainstream pop. A fairness analytic approach must distinguish between statistical bias in transcription accuracy and representational harm caused by treating certain traditions as outliers. Sociotechnical fairness research [20] warns against abstract fairness metrics that ignore institutional context. Music AI systems should therefore support community-defined evaluation and provide disaggregated performance reporting. The design of graph schemas should be informed by music theory and ethnomusicology, not only by engineering convenience.

Governance requires clear roles for data creators, rights holders, researchers, and platform operators. Music corpora contain copyright-sensitive audio and scores, and the inclusion of singing voice recordings raises privacy and consent issues. Governance frameworks should include data provenance, licensing, cultural consultation, and mechanisms for withdrawal. Graph models trained on these corpora inherit data governance decisions, especially when node features are derived from original recordings. A graph model is not neutral with respect to these decisions; its relational structure can expose or obscure attribution and source information.

Robustness in musical settings also includes resilience to distribution shift. A model trained on clean multitrack stems may fail on historical recordings or user-generated video. Evaluation should cover segmentation stability under noise, tempo changes, transposition, and partial occlusion of vocals. Continuous monitoring and benchmark updates are essential. Governance mechanisms should require documentation of intended use and known failure modes. This documentation should be accessible to music archivists, educators, and performers, not only to machine learning specialists.

7. Deployment, Sustainability, and Policy Implications

Sustainability and policy concerns are increasingly relevant to music AI infrastructures. Environmental analyses of large AI systems [21] suggest that training and inference energy use should be a design constraint, not an afterthought. Graph neural network music models can be made more sustainable through node pooling, edge sparsification, transfer learning, and selective re-encoding. However, efficiency gains must be balanced against representational richness. Policy incentives can encourage reporting of computational costs and support shared benchmark hubs. Shared evaluation infrastructure reduces redundant training and lowers the environmental footprint of the research community.

Broader policy implications from data-intensive systems also apply to music platforms [22]. Surveillance-like collection of listening and performance data can shape creative practice and user autonomy. Systems that analyze singing and melody may be deployed in educational personalized feedback, copyright enforcement, recommendation, and archival description. Each deployment context has different error costs and governance requirements. Copyright enforcement may demand high precision, whereas educational feedback may tolerate uncertainty. Policymakers and system designers should align technical performance criteria with these context-specific values.

Looking forward, cross-institutional datasets and model cards for music structure systems should document graph schemas, training data diversity, evaluation splits, and energy consumption. Community standards can reduce fragmentation and improve comparability. Integration with music theory scholarship and ethnomusicology can enrich node and edge definitions, making graph-based music analysis more inclusive and interpretable. Long-term sustainability will depend on building institutions that maintain shared resources, support cultural consultation, and create incentives for transparent and responsible system development.

8. Conclusion

Graph neural networks provide a powerful framework for representing musical structure and singing-melody relationships as relational systems rather than flat sequences or isolated features. This paper has argued that the design of such systems involves more than model architecture. Node granularity, edge semantics, graph construction, data provenance,

evaluation strategies, and deployment contexts all shape how musical knowledge is encoded and how the system behaves in practice. Architectural modularity and scalability are necessary for large-scale use, but they must be accompanied by robustness monitoring, fairness analysis, and governance mechanisms that respect the diversity of musical traditions and the rights of performers.

The systems perspective developed here emphasizes that graph-based music analysis tools operate as socio-technical infrastructures. They affect archival description, music education, creative analysis, and platform recommendation. Their design should therefore be informed by interdisciplinary expertise and by the communities whose music they process. Future work should develop transparent evaluation protocols, spatially and temporally efficient graph models, community-centered dataset governance, and policy frameworks that connect technical choices to broader social and environmental outcomes. By integrating graph-based relational modeling with careful systems design, researchers can build music analysis systems that are not only accurate but also accountable, inclusive, and sustainable.

References

1. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*.
2. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. *International Conference on Learning Representations*.
3. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), 4–24.
4. McFee, B., & Ellis, D. P. W. (2014). Analyzing song structure with spectral clustering. In *Proceedings of the 15th International Society for Music Information Retrieval Conference* (pp. 405–410).
5. Serrà, J., Müller, M., Grosche, P., & Arcos, J. L. (2014). Unsupervised music structure annotation by time series structure features and segment similarity. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(7), 1229–1240.
6. Zhang, Y., Pan, C., Guo, W., Li, R., Zhu, Z., Wang, J., ... & Zhao, Z. (2024). Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. *Advances in Neural Information Processing Systems*, 37, 1117-1140.
7. Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., ... & Eck, D. (2019). Music transformer: Generating music with long-term structure. *International Conference on Learning Representations*.
8. Kim, J. W., Salamon, J., Li, P., & Bello, J. P. (2018). CREPE: A convolutional representation for pitch estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 161–165).
9. Bittner, R. M., McFee, B., Salamon, J., Li, P., & Bello, J. P. (2017). Deep salience representations for F0 estimation in polyphonic music. In *Proceedings of the 18th International Society for Music Information Retrieval Conference* (pp. 63–70).
10. Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., & Ellis, D. P. W. (2014). *Mir_eval*: A transparent implementation of common MIR metrics. In

Proceedings of the 15th International Society for Music Information Retrieval Conference.

11. Müller, M. (2015). *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer.
12. Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4), 668–696.
13. Nieto, O., & Bello, J. P. (2016). Systematic exploration of computational music structure research. In *Proceedings of the 17th International Society for Music Information Retrieval Conference*.
14. Jeong, D., Kwon, T., Kim, Y., & Nam, J. (2019). Graph neural network for music score data and its application to emotion prediction. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*.
15. Paulus, J., Müller, M., & Klapuri, A. (2010). Audio-based music structure analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference* (pp. 625–630).
16. Stoller, D., Ewert, S., & Dixon, S. (2018). Wave-U-Net: A multi-scale neural network for end-to-end audio source separation. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*.
17. Meseguer-Brocal, G., & Peeters, G. (2019). Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separation. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*.
18. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
19. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).
20. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68).
21. Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
22. Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.