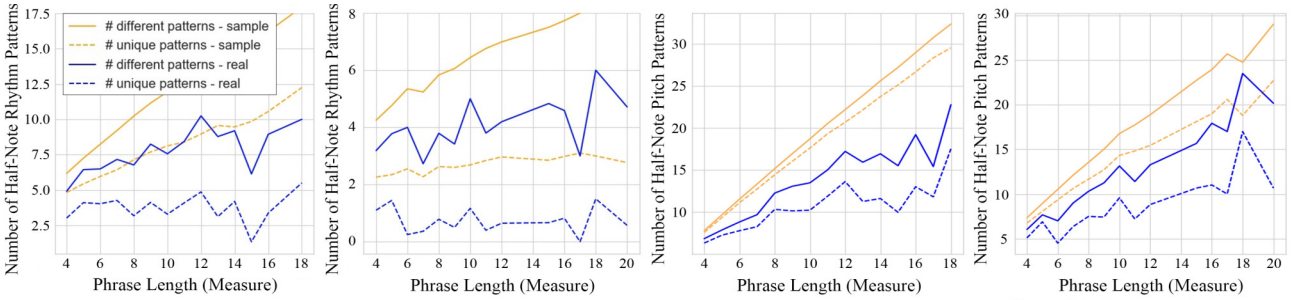


(a) POP909 phrases – rhythm pattern (b) PDSA phrases – rhythm pattern (c) POP909 phrases – pitch pattern (d) PDSA phrases – pitch pattern

Figure 3: Average number of different patterns (solid) and unique patterns (dashed) within a phrase vs. phrase length. (a)(b) are patterns of half-note melody note onset, (c)(d) are half-note pitch patterns. Blue lines are real phrases in the dataset. Orange lines are sampled phrases constructed by choosing each pattern at random from the entire dataset distribution.

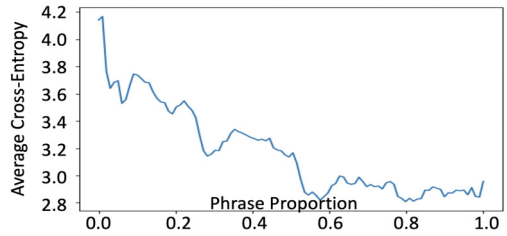


Figure 4: Average cross-entropy of diatonic pitch predictions over time within a phrase in POP909, using variable-order Markov models.

We extract the phrase-level structure in PDSA using the algorithm in [13], and use human-labeled structure in [13] for POP909. We study rhythm patterns by segmenting melody sequences into half-note onset patterns. PDSA has 54 different half-note patterns, while POP909 has all 128 possible patterns (onsets are quantized to 16th notes, and we assume an initial onset).

We claim that phrases have a limited vocabulary of onset patterns and therefore much repetition. This will of course be true necessarily if the entire dataset has a very limited vocabulary, so as a baseline for comparison, we construct random phrases by sampling half-note onset patterns from the entire dataset distribution. For POP909 songs, Figure 3(a) shows the average number of different onset patterns in phrases of different lengths (solid lines) and also the average number of patterns that occur only once in the phrase (dashed lines). This allows us to evaluate whether patterns within phrases (blue) have similar distributions to those of the entire dataset (orange). A similar graph for PDSA songs is shown in Figure 3(b). The analysis of pitch patterns is similar. Figure 3(c)(d) presents counts of melodic pitch patterns analogous to the onset pattern counts. Again, we see that real song phrases have a limited vocabulary compared to phrases assembled from random half-note units representing the entire dataset, and fewer real phrases go unrepeated.

If repetition of small patterns is frequent, variable-order Markov models [38, 39] can make good song-specific online predictions by updating the model at each element of the observed sequence. Conceptually, a variable-order Markov model estimates separate Markov chains of order 0 to N based on observed state (e.g., pitch) transitions. When data is too sparse to use a higher order, the model falls back (iteratively) to the next lower order. Repetitions of different lengths are readily learned, thus prediction accuracy indicates repetition significance. Figure 4 shows the average cross-entropy for pitch prediction over the length of all phrases, with phrase length normalized to 1. The clear downward trend shows the extent to which internal repetition enables better prediction.

In summary, we find abundant evidence for repetition within phrases:

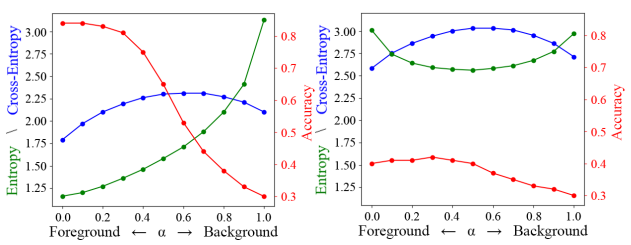
- Compared to sampling onset patterns at random from the POP909 distribution, real phrases have fewer distinct onset patterns, and more onset patterns are repeated in the phrase (Fig 3(a)). Same trend occurs in PDSA (Fig 3(b)), whose phrases contain even fewer distinct onset patterns.
- Rhythm patterns show a clear repetition structure within phrases. Considering each measure as a rhythm pattern, in PDSA, 7% of phrases have the same rhythm pattern in every measure, 28% have a repetition structure in one of the forms: "abab," "aabbaabb," "aba," or "abababa."
- The vocabulary of pitch patterns within a song or phrase is also very limited compared to the whole dataset, implying pitch sequence repetitions within the phrase level.
- The cross-entropy of predicting pitches decreases over time in a phrase, suggesting within-phrase repetition.

3.1.3 Song-Specific Vocabulary and Common Patterns

Motives and patterns below the phrase level create a song-specific vocabulary and have a long-term dependency throughout the whole piece. For example, in Beethoven's Symphony No.5, the first movement begins with a distinctive four-note "short-short-short-long" (the famous "Fate Motive"). It repeats everywhere throughout the piece, and Beethoven arranged it logically to unify the entire work.

We chose variable-order Markov models to further explore patterns because (a) it is easy to tune the weights between models trained on different data, (b) models can learn different lengths of pattern repetition. The following results show 1) song-specific vocabulary is critical in prediction, 2) there is long-term dependency between phrases.

To see the effects of song-specific vocabulary and context, we compare the results of variable-order Markov models predicting pitch trained on many other songs (background model), versus training on the same song after removing a short segment to be predicted (foreground model). Predictions are a linear combination of the back-



(a) Including duplicated phrases in foreground training

(b) Hold out repeated phrases in foreground training

Figure 5: Average entropy, cross-entropy and accuracy of predicting the first eight diatonic pitches in each phrase, tuning between foreground and background in POP909.

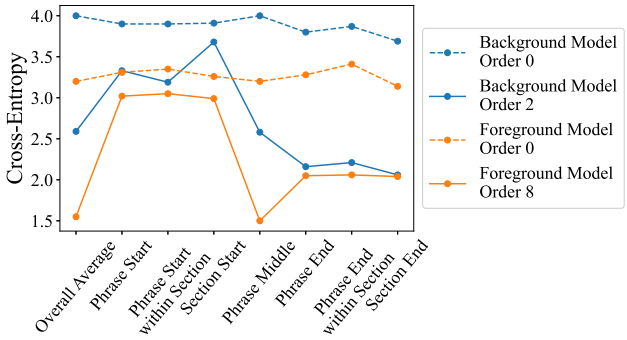


Figure 6: Average cross-entropy on diatonic pitches at different structure level positions in POP909 dataset predicted by background and foreground variable Markov models.

ground and foreground models. Figure 5(a) shows that the pure foreground model always obtains the best results. One possible reason is that phrase-level repetition in the foreground allows memorization. Thus, we removed all the duplicated or similar phrases in the foreground training and tested only on the first eight notes in each phrase (without updating the model) to eliminate both the effects of phrase-level repetitions and within-phrase motive repetitions. With almost no information from the repeated phrase, we might expect little information within a song to help predict pitches. However, Figure 5(b) indicates that we obtain the best prediction accuracy when incorporating 70% foreground and 30% background information. Since duplicate phrases have been removed and the first eight notes involve little within-phrase repetition, the results show that there must be a song-specific vocabulary or context that repeats across different phrases throughout the song, but not across the database (otherwise background would work better).

3.2 Structural Influence

Can repetition structure be considered orthogonal to melody, harmony and rhythm generation? If so, we can simply generate music note-by-note, ignoring structure, and then impose structure by repeating generated sequences. However, if there is significant interaction between structure and other facets of music, then structure is an integral part of music analysis, music modeling, and music generation. Our findings show the latter is the case.

Previous work has found systematic interactions between repetition structure and three facets of music: melody, rhythm and harmony, with interactions at both the phrase level and the section level [13]. For example, chords at the ends of phrases differ from chords in the middle of phrases. Furthermore, final chords in sections differ significantly from final chords in phrases that are not at the ends of sections. This does not mean that "good music" must reflect a structural hierarchy, but at least this finding offers insight into how music generation might be improved, and it raises questions for further study.

In this work, we have also explored confidence and surprise in music construction to better understand the role of different levels of structure in music. In Figure 6, we trained on the entire dataset (background model) of melody pitch sequences, holding out test songs; and also trained on single songs (foreground model), holding out all phrase repetitions to eliminate prediction by memorizing phrases, and holding out eight notes at a time for testing. Only order 0 (histogram) and the best-performing order maxima (2 and 8, respectively) are shown. As seen in the previous section, the foreground model predictions are better by 1 bit on average, indicating more repetition within songs (foreground) than across songs (background).

Even though the Markov models have no positional encoding or conditions, we can still see a dramatic difference in predictability at different structure positions (Figure 6). For example, using the background model, cross-entropy at the start of sections is over 3.5 (bits), while other phrase starts are about 3.2, and for most notes (phrase middle), cross-entropy is only 2.6. Using the foreground model, we see a different pattern, but predictability still varies substantially according to position in the structure.

3.3 Repetition and Structure Over Time

Music is not repeated randomly. After seeing different levels of hierarchy in Section 3.1, we ask: Is there a schema for repetition? How does repetition play out over time?

Huron suggests that if music is to manipulate prediction through repetition, it makes sense to repeat some of the music early on [14]. This affords immediate pleasure from successful prediction rather than delaying until all novel material is exhausted. POP909 supports this hypothesis: More than 50% of the phrases repeat immediately, and almost all phrases repeat within a quarter of the song.

In Section 3.1, we have seen that most rhythmic and melodic patterns in a phrase are repetitions. At the song level as well, using the phrase repetition labels in POP909, we found that for 79% of songs, 15% to 35% of their duration consists of new material and the rest is repetition.

Returning to our consideration of structure over time: How does surprise vary with structure? We might expect less surprise at the ends of sections to give a sense of completion or resolution. Figure 7 Left shows a histogram percentage of phrase-level repetition over the course of a song. We see a relatively low repetition rate in the first 1/10 of the song. The repetition rate sharply increases as we progress to the first 1/5 of the song because, after introducing the initial materials, most songs repeat them. There is a notice-

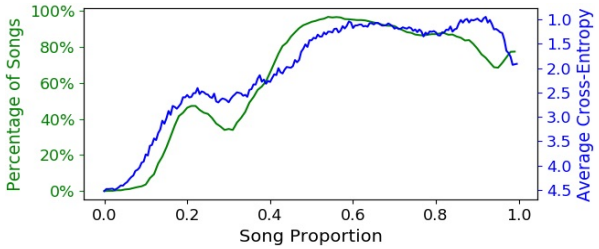


Figure 7: **Left** (green): Percentage of POP909 songs that have phrase repetitions at different song locations. **Right** (blue): Average prediction cross-entropy using variable-order Markov models on diatonic pitches in POP909 songs over time, and note the y-axis is inverted.

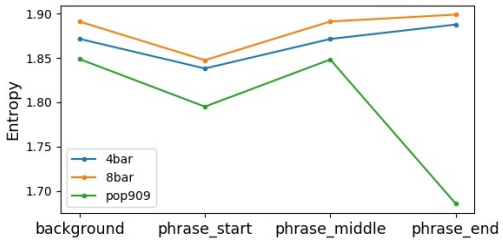


Figure 8: Entropy of pitch scale degree distribution at different phrase positions, comparing generated 4-bar and 8-bar phrases from VAE with real phrases in POP909.

able drop around the quarter-way point where many songs introduce new material. In the second half, almost everything is a repetition or variation of what came in the first half. Finally, the graph shows novel material is often introduced near the end. In Figure 7 Right, we use the variable-order Markov model to calculate the average cross-entropy over time on melody pitches. To show the correspondence, we flip the vertical cross-entropy axis. Note the similarity between the trends in repetition and cross-entropy.

From these results, it is clear that repetition is not random but follows a plan in which novelty is revealed, presumably to enhance the enjoyment or effect of the music. This is perhaps surprising because other research shows that music can be substantially rearranged without destroying positive impressions [40]. Whether this organization matters to listeners or simply reflects composers' intentions requires further research.

4. IMPLICATIONS FOR DEEP MUSIC GENERATION

One application of our studies is to gain insight into deep learning models for music generation. We can apply the analyses from Section 3 to melodies generated from deep learning models. Through these case studies, we can characterize repetition and structure from deep music generation and compare to human-composed songs. To be clear, we are not claiming that any particular structure is *necessary* or even *good*. Our goal is to illuminate possibilities and better understand both real and generated music. We then discuss our results and point out new directions and ideas for future work in deep music generation.

4.1 Lack of Repetition and Structure

We studied repetition and structure in deep learning generated melodies to answer three questions: 1) do melodies have multiple levels of repetition and structure? 2) do they have song-specific vocabulary and common patterns? 3) how does cross-entropy vary over time? We used two deep music generation models: One is a VAE model using representation learning [24], chosen because it uses contrastive learning to generate longer sequences (8 bars) than other VAE models. The other model is Music Transformer [11].

4.1.1 Do deep melodies have multiple levels of structure?

We want to know if deep-learning-generated music has multiple levels of structure and whether any structure is reflected in pitch and rhythm. Unfortunately, most deep learning models can generate only a limited length: VAEs and GANs usually have a fixed length from 2 to 8 bars, about the length of just one phrase. Sequential models like Transformer and LSTM can generate longer sequences, but we cannot find any salient repetition comparable to repeated phrases by listening or by automated analysis. These differences from popular songs are striking.

Thus, the only thing we can do is to test on short generated sequences, consider them as a phrase, and see if they have phrase structure. We tested the 4 bar and 8 bar generated phrases from a VAE model [24] and used 1000 examples of each duration. Figure 8 shows that there is no significant differences between the entropy of pitch scale degree distributions at the start of phrases, middle of phrases and end of phrases, but we see significant differences in the real POP909 phrases. Also, the probabilities of different pitches at different phrase-level positions only change slightly in the analyzed outputs. For example, the probability of seeing the tonic at phrase end is 35% in real POP909 phrases, while at the other positions it is around 20%. The probability of seeing the tonic in VAE output is in the range of 20% to 21% for all positions.

4.1.2 Do deep melodies have song-specific vocabulary and common patterns?

Following the logic in Figure 3, we count the rhythm (melody note onset) patterns for whole songs from Music Transformer and from POP909 itself. We trained the Music Transformer on 80% of songs in POP909, using 10% for validation and 10% for testing. We initialized the generation with the first 2 to 6 bars of the test songs, but did not count the initialization bars as part of the generated melody. To calculate the song length, we use the total length of melodic phrases, omitting measures that are empty or have long rests.

In Figure 9, we see that transformer-generated results have a much higher rhythm vocabulary compared to real songs. Many people have observed that deep-learning-generated music has many outlier notes and patterns that appear once and sound like mistakes. Here we see that transformer-generated melodies have a higher number of unique rhythm patterns compared to real songs, which might explain the sense of too many outliers.

We also analyzed half-note pitch patterns in 4-bar and 8-bar phrases from the PDSA, from random pitch patterns drawn from the entire PDSA distribution, and for the VAE

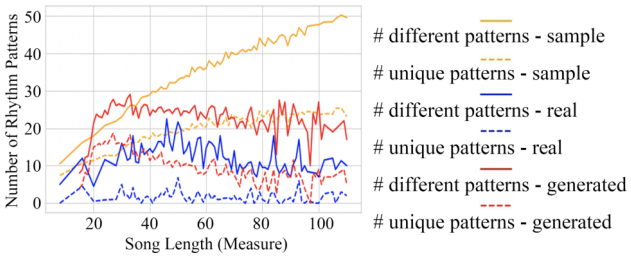


Figure 9: Analysis of note onset pattern vocabulary from randomly sampled POP909 patterns (yellow), Music Transformer songs (red), and POP909 songs (blue) as a function of song length. Solid lines are the total number of different patterns, and dashed lines are the number of unique (not repeated) patterns.

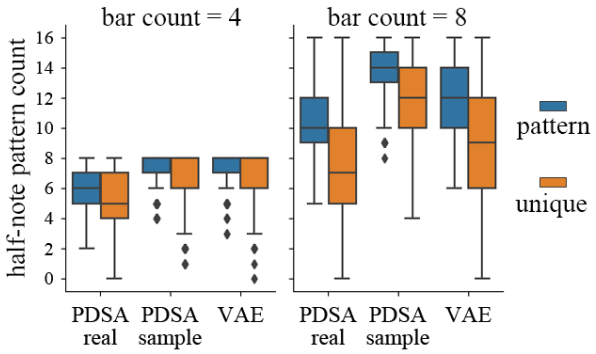


Figure 10: Comparison of half-note pitch pattern counts in PDSA phrases, randomly sampled PDSA pattern phrases, and phrases from VAE. The y-axis counts the number of different patterns (blue) or number of unique (not repeated) patterns (orange) found in 4 or 8 bars. The VAE generated patterns are significantly more than the patterns in the real PDSA dataset, with p-value less than 10^{-5}.

model [24] (see Figure 10). Vocabulary is more limited in PDSA phrases compared to the random patterns and VAE patterns. The VAE pattern counts are closer to those of synthetic phrases and thus may lack some of the coherence, redundancy, and predictability of PDSA.

4.1.3 Entropy over time

We also compared trends in cross-entropy of the pitch sequence over the course of songs from POP909 (Figure 7 blue) and Music Transformer (Figure 11). Although both exhibit an overall cross-entropy decrease over the course of the phrase, there are important differences. POP909 cross-entropy increases at around 20% of the song length, probably due to the introduction of novel and contrasting patterns, and also around the end. POP909's average cross-entropy is greater than one bit per note except for a small region around 90% of song length, while Music Transformer is below one on average for the last 30% of the song, suggesting overly predictable sequences.

4.2 Discussion and New Directions

Rather than learning and reproducing general statistics of datasets, we need to learn how songs strategically diverge from background or stylistic norms to create interest, surprise, and individuality. It is particularly interesting that

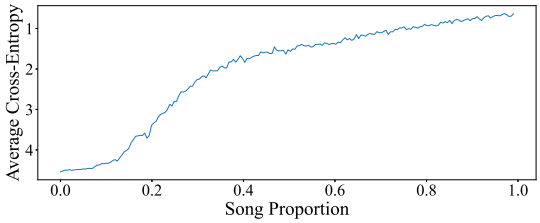


Figure 11: Average prediction cross-entropy using Markov models on diatonic pitches over time in 910 generated music pieces from music transformer, y-axis inverted.

phrases can be better predicted by relatively short phrases within the same song than by large amounts of training data from other songs. It seems plausible that songs of the same artist or same sub-genre may be more predictive than songs in general. Our findings also reinforce previous findings that using structure in MusicFrameworks [12] results in better human evaluations.

Examples in Section 4.1 suggest that we can compare generated music to real music using measures of structure, repetition and entropy. Matching these measures is not guaranteed to make music "better," but we should not simply ignore clear objective differences. We would at least expect differences to be small when the task is to imitate a style or genre. We can also speculate that these measures are relevant to listener preferences even if they do not tell a complete story.

Although we have focused on popular music, repetition and hierarchical structure seem to be ubiquitous in music. Pop music, with its nearly exact repetitions, seems easier to study than Classical music where we might expect more variation, development and modulation, which make repetition less obvious. We are encouraged by our results using variable-order Markov models on pitch sequences. These studies reveal much more information than we expected and do not rely on finding wholly duplicated measures, which we used to extract phrases and sections. Perhaps this approach will be of use in Classical and other music.

5. CONCLUSION

It should be no surprise that structure, repetition, pitch, rhythm, harmony and entropy are all strongly connected and interdependent. We have offered new ways to explore these connections objectively, using a data-driven approach without relying on subjective human analyses.

Among our findings are that within-song and within-phrase vocabulary and repetition are not a reflection of more general background statistics from a collection of songs. Instead, songs and phrases gain "individuality" through more repetition and smaller vocabulary. This has important implications for machine learning and music generation systems.

There are clear differences between measurements of real songs and those of many music generation systems, suggesting that there are important gaps to fill through new research. We hope that this work will inspire further research in the roles played by repetition and structure in music as well as methods to learn repetition and structure.

6. ACKNOWLEDGEMENT

We would like to thank Ziyue Piao and Ying Yee Wong for their assistance in the early stage.

7. REFERENCES

[1] W. E. Caplan, *Classical Form: A Theory of Formal Functions for the Instrumental Music of Haydn, Mozart, and Beethoven, Revised Edition*. Oxford University Press, 2000.

[2] P. Goetschius, *Lessons in music form: A manual of analysis of all the structural factors and designs employed in musical composition*. Library of Alexandria, 1904, vol. 1.

[3] D. M. Green, *Form in tonal music*. Holt, Rinehart and Winston, 1979.

[4] R. B. Dannenberg and M. Goto, "Music structure analysis from acoustic signals," *Handbook of Signal Processing in Acoustics*, vol. 1, pp. 305–331, 2009.

[5] P. Allegraud, L. Bigo, L. Feisthauer, M. Giraud, R. Groult, E. Leguy, and F. Levé, "Learning sonata form structure on mozartś string quartets," *Transactions of the Int. Society for Music Information Retrieval Conf.*, vol. 2, Dec. 2019.

[6] J. Salamon, "Deep embeddings and section fusion improve music segmentation," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[7] O. Nieto, G. J. Mysore, C. C. Wang, J. B. L. Smith, J. Schlüter, T. Grill, and B. McFee, "Audio-based music structure analysis: Current trends, open challenges, and applications," *Transactions of the Int. Society for Music Information Retrieval Conf.*, vol. 3, no. 1, p. 246–263, 2020.

[8] M. Hamanaka, K. Hirata, and S. Tojo, "Musical structural analysis database based on GTTM," in *Proc. of the 15th Int. Society for Music Information Retrieval Conf.*, 2014.

[9] E. N. G. Shibata, R. Nishikimi and K. Yoshii, "Statistical music structure analysis based on a homogeneity-, repetitiveness-, and regularity-aware hierarchical hidden semi-markov model," in *Proc. of the 20th Int. Society for Music Information Retrieval Conf.*, 2019.

[10] J. Paulus, M. Muller, and A. Klapuri, "Audio-based music structure analysis," in *Proc. of the 11th Int. Society for Music Information Retrieval Conf.*, 2010, pp. 625–636.

[11] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," *arXiv preprint arXiv:1809.04281*, 2018.

[12] S. Dai, Z. Jin, C. Gomes, and R. B. Dannenberg, "Controllable deep melody generation via hierarchical music structure representation," in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, 2021.

[13] S. Dai, H. Zhang, and R. B. Dannenberg, "Automatic analysis and influence of hierarchical structure on melody, rhythm and harmony in popular music," in *in Proc. of the 2020 Joint Conference on AI Music Creativity (CSMC-MuMe 2020)*, 2020.

[14] D. Huron, *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA: MIT Press, 2006.

[15] E. Narmour *et al.*, *The analysis and cognition of melodic complexity: The implication-realization model*. University of Chicago Press, 1992.

[16] E. H. Margulis, *On Repeat: How Music Plays the Mind*. Oxford University Press, 2013.

[17] J. J. Rolison and J. Edworthy, "The role of formal structure in liking for popular music," *Music Perception: An Interdisciplinary Journal*, vol. 29, no. 3, pp. 269–284, 2012.

[18] O. Julian and C. Levaux, Eds., *Over and Over: Exploring Repetition in Popular Music*. Bloomsbury Academic, 2018.

[19] J. Summach, "The structure, function, and genesis of the prechorus," *Music Theory Online*, vol. 17, no. 3, October 2011.

[20] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck, "A hierarchical latent vector model for learning long-term structure in music," in *Proc. of the International conference on machine learning*. PMLR, 2018, pp. 4364–4373.

[21] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[22] C. Payne, "Musenet," *OpenAI, openai.com/blog/musenet*, 2019.

[23] J.-P. Briot, G. Hadjeres, and F. Pachet, "Deep learning techniques for music generation," *Springer*, vol. 10, 2019.

[24] S. Wei and G. Xia, "Learning long-term music representations via hierarchical contextual constraints," in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, 2021.

[25] L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[26] G. Hadjeres and L. Crestel, "Vector quantized contrastive predictive coding for template-based music generation," *arXiv preprint arXiv:2004.10120*, 2020.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[28] G. Medeot, S. Cherla, K. Kosta, M. McVicar, S. Abdallah, M. Selvi, E. Newton-Rex, and K. Webster, "Structurenet: Inducing structure in generated melodies." in *Proc. of 19st Int. Conference on Music Information Retrieval Conf., ISMIR*, 2018, pp. 725–731.

[29] T. Collins and R. Laney, "Computer-generated stylistic compositions with long-term repetitive and phrasal structure," *Journal of Creative Music Systems*, vol. 1, no. 2, 2017.

[30] C.-Z. A. Huang, T. Cooijmans, A. Roberts, A. Courville, and D. Eck, "Counterpoint by convolution," in *Proc. of the 18th Int. Society for Music Information Retrieval Conf.*, Suzhou, China, 2017.

[31] C.-H. Chuan and D. Herremans, "Modeling temporal tonal relations in polyphonic music through deep networks with a novel image-based representation," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[32] L.-C. Yang and A. Lerch, "On the evaluation of generative models in music," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4773–4784, 2020.

[33] A. Elowsson and A. Friberg, "Algorithmic composition of popular music," in *Proc. of the 12th Int. Conference on Music Perception and Cognition and the 8th Triennial Conf. of the European Society for the Cognitive Sciences of Music*, 2012, pp. 276–285.

[34] S. Dai, X. Ma, Y. Wang, and R. B. Dannenberg, "Personalized popular music generation using imitation and structure," *arXiv preprint arXiv:2105.04709*, 2021.

[35] B. L. Sturm and O. Ben-Tal, "Taking the models back to music practice: Evaluating generative transcription models built using deep learning," *Journal of Creative Music Systems*, vol. 2, pp. 32–60, 2017.

[36] Z. Wang, K. Chen, J. Jiang, Y. Zhang, M. Xu, S. Dai, G. Bin, and G. Xia, "Pop909: A pop-song dataset for music arrangement generation," in *Proc. of 21st Int. Conference on Music Information Retrieval Conf.*, 2020.

[37] D. Berger and C. Israels, *The Public Domain Song Anthology*. Charlottesville: Aperio, Mar 2020.

[38] R. Begleiter, R. El-Yaniv, and G. Yona, "On prediction using variable order markov models," *Journal of Artificial Intelligence Research*, vol. 22, pp. 385–421, 2004.

[39] J. Cleary and I. Witten, "Data compression using adaptive coding and partial string matching," *IEEE transactions on Communications*, vol. 32, no. 4, pp. 396–402, 1984.

[40] Z. Eitan and R. Y. Granot, "Growing oranges on mozart's apple tree: "inner form" and aesthetic judgment," *Music Perception: An Interdisciplinary Journal*, vol. 25, no. 5, pp. 397–417, 2008.

HETEROGENEOUS GRAPH NEURAL NETWORK FOR MUSIC EMOTION RECOGNITION

Angelo Cesar Mendes da Silva
Universidade de São Paulo, Brazil
angelo.mendes@usp.br

Diego Furtado Silva
Universidade Federal de São Carlos, Brazil
diegofs@ufscar.br

Ricardo Marcondes Marcacini
Universidade de São Paulo, Brazil
ricardo.marcacini@usp.br

ABSTRACT

Music emotion recognition has been a growing field of research motivated by the wealth of information that these labels express. Recognition of emotions highlights music's social and psychological functions, extending traditional applications such as style recognition or content similarity. Once musical data are intrinsically multi-modal, exploring this characteristic is usually beneficial. However, building a structure that incorporates different modalities in a unique space to represent the songs is challenging. Integrating information from related instances by learning heterogeneous graph-based representations has achieved state-of-the-art results in multiple tasks. This paper proposes structuring musical features over a heterogeneous network and learning a multi-modal representation using Graph Convolutional Networks with features extracted from audio and lyrics as inputs to handle the music emotion recognition tasks. We show that the proposed learning approach resulted in a representation with greater power to discriminate emotion labels. Moreover, our heterogeneous graph neural network classifier outperforms related works for music emotion recognition.

1. INTRODUCTION

Music is intrinsically connected to human emotions. At the same time that a songwriter expresses emotions in the composed songs, we can use music's emotional information to characterize the listener's moments and feelings. Music emotion recognition (MER) is a task that highlights social and psychological functions comprised by recordings, extending traditional applications in the area of music information retrieval [1]. According to the literature, mapping the spectrum of emotions can be done by analyzing the values in arousal and valence domains [2]. Besides, the emotions can be represented by labels that indicate negative, positive, and neutral values of arousal and valence and labels that point to emotional expressions, such as happiness, anger, or satisfaction [3]. Each label may be defined according to the significance of valence and arousal simultaneously, as illustrated in Figure 1. The MER task aims to use a musical representation to estimate these significances or predict discretized labels.

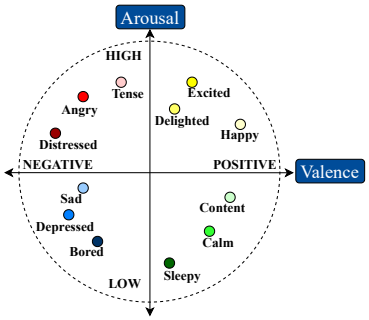


Figure 1: Emotions mapped in the arousal and valence domains.

Textual information is the most used data type in tasks involving emotion labels [4]. When emotion recognition is performed on music data, lyrics are commonly used as input [5, 6]. However, some studies emphasize the importance of acoustic features for the same task [7]. The fact that both modalities complement musical perception is well-known [8–10]. However, there is no consensual approach for incorporating multiple modalities into a unified representation.

Defining the structure to deal with different data types, such as text and audio, is a preliminary step to generate a unified multi-modal representation of musical data. Building this structure is an open problem in the literature. Techniques based on feature fusion are the most explored approach in the literature [11, 12]. Fusion of features can be performed by simply concatenating text and audio features or even learning embeddings via multimodal deep learning. However, these existing methods have limitations regarding incomplete modalities, deal with modalities of different dimensionality, and the interpretability of the generated representations [13].

Heterogeneous networks are a well-known representation for manipulating multiple modalities of unstructured data [14, 15]. This resource has been explored due to the ability to map data on graphs, representing connections between objects (nodes) through edges. Therefore, this process produces a graph that can be used as input in Graph

Neural Network models [16]. Structuring data on heterogeneous networks has been widely studied in varied data such as text and images [17,18], but there are still few studies for musical data.

This work introduces a novel model for music emotion recognition that uses a heterogeneous network to structure audio and lyrics features and build a new multi-modal graph-based representation of musical data using the graph convolutional network (GCN). Our proposed model, described in detail in Sec. 3, uses content-based features (audio descriptors) together with features extracted from lyrics (embeddings from the language model). The music graph topology is based on relations defined by cluster information from audio e textual metadata. Some music nodes have labels with emotional information. Therefore, the proposed structure is composed of a heterogeneous network, and GCN can predict the emotion of unlabeled music nodes.

We apply a network regularization framework to propagate and refine information between neighboring nodes from different modalities. The matrix resulting from this regularization, along with the graph connections, is used as input for training our GCN. In this embedding space, music with similar emotions are close to each other, while dissimilar ones are further apart.

To evaluate our approach, we used the publicly available dataset PMEmo [19], which has 794 songs represented by audio descriptors and lyrics with annotations for arousal and valence domains. We measured the relevance of the learning process of a multi-modal graph-based representation for MER. Our proposal was compared with two different methods and other works from the literature that reproduced a similar experiment.

Our method was able to obtain a meaningful embedding representation that unifies different audio and text music features across the heterogeneous network. In addition to the heterogeneous network enabling interpretability of the relationships between text and audio features, our method outperforms other existing methods for music emotion recognition tasks.

2. RELATED WORK

The construction of methods to learn representations for unstructured data aims to reduce the work for features extraction and selection to describe the data and make learning algorithms less dependent on these processes [20]. Learning a representation for musical data involves manipulating features obtained by multiple modalities and projected in different spaces [21, 22].

Learning to represent music to recognize the comprised emotion involves exploring different features [1, 23]. We can observe most applications using the lyrics to represent the music [5, 6, 24] justified by the direct association we can make with the meaning of words and emotion labels. However, we have evidence that acoustic features also contain information that characterizes emotions [11,12,25]. In summary, there is no consensus on representing music using multiple modalities for the MER task, although we note that audio and lyrics features are relevant.

Recent work about music representation learning has explored strategies based on deep neural networks, such as early and late fusion [26, 27] and evaluating combinations between different input features [13]. This strategy advances and presents successful results that justify the representation learning process. However, there is a demand for approaches that explain the semantic complementarity existing between the features that justify the obtained results [28]. We explore the heterogeneous network resources to embed multiples music features and explicit the latent relationships among nodes, similar to [29, 30]. Moreover, we learn a graph-based representation that concentrates information from related music.

The complementarity between distinct features in musical perception justifies the motivation to build a multimodal representation [31]. For the emotion recognition problem, [7] describes what semantic information exists in various acoustic features and which emotions each feature can emphasize. We note that lyrics are also associated with emotions and must be used for the recognition task [32]. Therefore, we aim to build a multi-modal representation that integrates features obtained through the audio and lyrics and compares its performance against unimodal scenarios.

The construction of multi-modal embeddings requires that distinct features be mapped in a unified space. Graph-based methods have been applied in recent years when nodes and edges have different types [33]. We can exploit the relationships between neighboring nodes to build an embedding space that aggregates information from the node features even in different spaces. Significant advances are observed in multi-modal and unstructured data applications [34].

For musical data, [35] explores the task of similarity between artists. Authors use data previously structured to create the graph and apply it in GCN to build music embeddings with acoustic characteristics and tags. Finally, the embeddings are used to predict links to new artists. In contrast, [36] introduces a proposal of graph-based representation learning with the graph being constructed from a co-occurrence matrix obtained by the presence of pairs of songs in playlists.

Our proposal aims to create a heterogeneous network with multiple musical features and connect nodes that share cluster information. This network produces the features input for GCN to build a multi-modal representation and handle the music emotion recognition task. Our learned representation does not depend on previous information to structure the graph. Although we do not extend the approach to the multi-task scenario, it can be easily adapted for different tasks, maintaining the same heterogeneous network.

3. MODELLING

Our goal is to recognize emotions present in songs, such as a classification problem $Y = f(X)$, where Y represents the set of emotion labels, X represents the songs, and $f(.)$