

RPPNET: PERCEPTUALLY-GROUPED RHYTHM-PITCH PRIMITIVES FOR LONG-TERM STRUCTURE MELODY GENERATION VIA BOUNDARY-AWARE MODELING

Tieyao Zhang¹ Yuke Liu¹ Jiaxing Yu²
Xinda Wu² Kejun Zhang² Genfang Chen^{1,3}

¹ Zhejiang Conservatory of Music, Hangzhou, China

² Zhejiang University, Hangzhou, China

³ ZJU-ZJCM MusicAI Lab, Hangzhou, China

tieyaozhang@stu.zjcm.edu.cn, keke@stu.zjcm.edu.cn, yujx@zju.edu.cn,
wuxinda@zju.edu.cn, zhangkejun@zju.edu.cn, cgf@zjcm.edu.cn

ABSTRACT

Existing symbolic music generation models typically use bars as the basic structural unit. However, human perception of musical phrases often does not align with notated bar lines, leading to long-term structural fragmentation. This paper proposes RPPNet—a two-stage deep learning architecture with variable structural boundaries. It first generates variable-length Rhythm-Pitch Primitive (RPP) sequences, where each RPP encodes note count, rhythm, and contour; then decodes the RPP sequences into concrete notes. The grouping of RPPs is automatically derived from acoustic cues, auditory inertia, and similarity perception based on music psychology. Experiments show that melodies generated by RPPNet are superior in both long-term structure and musicality, with significant improvements across all subjective evaluation dimensions. Ablation studies confirm that the performance gain stems from the structural correctness of the psychological representation, rather than from model capacity. This work offers an interdisciplinary perspective for music generation, integrating music theory, computational modeling, and music psychology.

1. INTRODUCTION

Symbolic melody generation is a core direction of generative AI in art [1]. In recent years, text-to-audio models have achieved notable results and can generate complete songs, yet they do not allow fine-grained user control [2–4]. Symbolic music, in contrast, enables precise control over melody, harmony, and other elements, holding a central position in human-AI co-creation. Consequently, this direction continues to attract significant attention from both academia and industry [5].

In recent years, researchers have drawn on language models from natural language processing to capture long-term dependencies in music and achieve structured melody generation. Such methods offer an end-to-end framework, excellent representation learning ability, and the potential to generate sequences of arbitrary length.

They require no hand-crafted domain rules and can automatically extract knowledge from large-scale data; their effectiveness has been validated across multiple disciplines [6–10].

However, long-term structure modeling in symbolic music remains challenging [11–13]. Existing methods operate within linear sequence frameworks where attention mechanisms capture long-range token correlations yet fail to learn explicit associations between non-adjacent but functionally similar structural units; meanwhile, this task requires integrating music theory, computational modeling, psychology, and neuroscience [14]. Both factors jointly constrain effective learning of hierarchical logic, biasing models toward local statistical patterns and limiting long-term structure generation quality.

Recent works integrate computational modeling with music theory to improve long-term coherence through explicit structural boundaries and thematic features, such as Museformer [15] with its hierarchical attention mechanism, Theme Transformer [16] with thematic material retrieval, and MELONS [17] and PopMNet [18] with graph-structured thematic variation encoding. However, these methods rely on fixed bar-level units. Psychological evidence shows that listeners perceive structural boundaries through acoustic cues, auditory inertia, and similarity-based grouping [19–21], rather than fixed metric segmentation. This mismatch hinders listeners’ perceptual grouping of the generated structure, thereby reducing overall musical coherence.

To address this problem, we propose a hierarchical melody generation framework, RPPNet¹. It is based on an autoregressive architecture and splits generation into an RPP-level generation stage and a Note-level generation stage. Both stages use a Transformer encoder-decoder [22].

Drawing on music theory and cognitive psychology, we use three perceptual dimensions—acoustic cues, auditory inertia, and similarity relations—and rhythmic structure analysis [20, 23, 24] to design a heuristic algorithm.

¹ The source code is available at: <https://github.com/Kreuzter0421/RPPNet-PyTorch.git>

This algorithm represents a melody as a sequence of non-equidistant rhythm-pitch primitives (RPPs).

The RPP-level stage employs decoupled serial prediction for autoregressive RPP generation, constructing long-term structure; the Note-level stage decodes RPPs into MIDI notes via time-scale expansion mapping. Unlike bar-based approaches (e.g., Museformer, MELONS), RPPNet replaces fixed bars with perceptually-driven flexible grouping units—the first hierarchical model to do so, grounded in psychological evidence that phrase boundaries often misalign with notated bar lines.

Our contributions are as follows:

1. We propose RPPNet, a two-stage hierarchical framework with a novel perception-driven structural representation. It replaces fixed bar-level boundaries with flexible rhythm-pitch primitives grouped by acoustic cues, auditory inertia, and similarity perception.
2. We propose an automatic grouping algorithm that derives structural boundaries without explicit annotations. Controlled ablations demonstrate that performance gains stem from perceptual grouping principles rather than model capacity.
3. The proposed time-scale expansion mapping bridges discrete RPP events and continuous Note-level generation, with decoupled serial prediction ensuring strict alignment between generated notes and the RPP sequences.
4. Experimental results show that RPPNet outperforms existing baselines in long-term structural coherence and musicality.

2. RELATED WORK

2.1 Long Sequence Symbolic Music Generation

Symbolic music generation performs algorithmic composition by modeling discrete musical events, encoded in formats including MIDI, MusicXML, and ABC notation. Early research focused on local sequential dependency modeling: N-gram Markov chains predict the next event based on joint preceding events [25], yet can only capture short-range dependencies and fail to model deep hierarchical structures. HMM introduce latent states to implicitly control token distributions [26, 27], while dynamic Bayesian networks further explicitly model deep dependencies [28].

In recent years, deep neural network generative models have achieved significant advances in symbolic music. From RNNs [29], VAEs [30], and GANs [31], to Transformers [32–34], modeling capabilities have steadily increased. However, the quadratic complexity of Transformer self-attention, compounded by the long sequence nature of symbolic music, poses severe challenges for modeling dependencies spanning hundreds of bars.

This challenge stems from standard tokenization strategies, where a single piece can reach thousands to tens of

thousands of tokens [35], severely weakening long-term dependency modeling [36, 37]. Even when subsequent studies introduce compound tokens (e.g., Compound Word [38], Octuple [39]) to compress sequence length, models still struggle to capture thematic repetition and variation spanning dozens of bars—the core difficulty of long-term structure modeling.

2.2 Structure modeling guided by music theory

To address the challenges of long-range structural modeling in music, researchers have attempted to incorporate music theory priors into model design to guide globally coherent generation. The relevant methods can be grouped into three categories.

Structure-aware attention: Optimizing attention distribution to efficiently extract musical structural features. For example, Museformer [15] combines fine- and coarse-grained attention to capture bar-level structure; HAT [40] builds a harmony-aware hierarchical Transformer, using harmonic progressions to guide long-range dependency modeling.

External structural prior injection: Explicitly introducing boundary or thematic features to constrain the generation process. Ref. [41] uses segmentation modules and multi-scale attention to inject structural boundary information, aiding long-form generation; Theme Transformer [16] takes retrieved thematic material as a conditional sequence to enhance cross-segment thematic consistency.

Hierarchical modeling and task decomposition: break long sequence generation into multi-level subtasks for globally controllable structure. MusicFramework [42] captures repetitive patterns via hierarchical representations to support multi-level abstract melody generation; TM-CGAN [43] explicitly models themes with a VAE-CGAN hybrid architecture, balancing global coherence and local variation; PhraseLDM [36] performs non-autoregressive generation in a phrase-level latent space to yield 128-bar multi-track music, alleviating long-term structure fragmentation. PopMNet [18] models internal dependencies via chord-melody graphs, while MELONS [17] combines bar-level relational graphs with GraphRNN and Transformer for long-term structure modeling.

However, these methods all lack perceptual grounding in structural segmentation. Neither explicit bar-level splitting nor implicit attention-based partitioning can capture the diversity and perceptual uncertainty of phrase boundaries, leading to structural degradation in long-term generation. Developing perceptually grounded, compact structural units for efficient structure-aware long-term structure modeling remains a core challenge.

2.3 Perceptual Theories of Musical Structure

The challenge of modeling long-range structure in symbolic music [11–13] stems from the highly structured intrinsic nature of music: its elements are organized according to certain structural frameworks and expressed in human-understandable ways [14]. This framework exhibits significant multi-dimensional characteristics [44–

46], encompassing not only basic components such as harmony and phrases but also interweaving cognitive organizational principles like similarity perception [21], making the analysis and modeling of its internal logic highly complex [47].

Music psychology research shows that structural perception relies primarily on three mechanisms: (1) acoustic cues—using changes in intensity, pitch, duration, timbre, etc. to mark structural boundaries [20]; (2) auditory inertia—based on the establishment and fulfillment of expectations for musical events [19]; (3) similarity perception—associating structural elements through motivic repetition and variation [21]. Together, these mechanisms reveal the hierarchical and theme-driven nature of musical structure.

Deliege [20] validated the psychological effectiveness of grouping preference rules, confirming that acoustic changes in intensity and pitch serve as core cues for perceiving structural boundaries. Cooper and Meyer [19] identified auditory inertia as the organizing force of musical rhythm: listeners perceive continuous sound as a dynamic structure based on psychological expectations, where deviations such as syncopation and rests create tension. Lamont and Dibben [21] demonstrated through cross-genre listening experiments that similarity judgments rely primarily on surface features such as texture and pitch contour rather than deep structural assumptions, exhibiting significant contextual dependence.

Traditional music modeling is mostly based on note- or bar-level granularity, which hardly aligns with the actual mechanisms of structural perception. Drawing on the above psychological findings, this paper proposes a heuristic algorithm that automatically analyzes monophonic melody structure based on acoustic cues, auditory inertia, and similarity perception, and converts it into RPP sequence data.

3. METHOD

3.1 Heuristic Segmentation of Rhythm-Pitch Primitive

To construct cognitively grounded low-level structural units, we propose a heuristic melody segmentation algorithm that parses melodies into Rhythm-Pitch Primitives (RPPs), each containing at most three notes to carry local boundary information and support RPP-level generation, laying the foundation for subsequent hierarchical modeling.

Primitives are defined by rhythmic patterns (five categories, e.g., iambic) and pitch contours (ascending, descending, etc.), as illustrated in Figure 1; evolutionary relationships between adjacent primitives are encoded via multi-hot vectors. Grouping employs dynamic programming to maximize the cumulative structural weight $W(\cdot)$ ² of the parsed sequence (Algorithm 1), where $W(\cdot)$ scores

² The complete formalization of $W(\cdot)$ is available in the repository mentioned above.

Algorithm 1 Dynamic Programming for RPP Grouping

Input: Note sequence $P_{1:L}$, Weight function $W(\cdot)$

Output: Optimal RPP list $\mathcal{R}$

```

1: if  $L \leq 3$  then
2:   return  $\{\text{MakeRPP}(P_{1:L})\}$ 
3: end if
4:  $dp[0] \leftarrow 0, \quad dp[1 \dots L] \leftarrow -\infty$ 
5:  $path[0 \dots L] \leftarrow \emptyset$ 
6: for  $i = 1$  to  $L$  do
7:    $k^* \leftarrow \arg \max_{1 \leq k \leq 3} \{dp[i-k] + W(P_{i-k+1:i})\}$ 
8:    $dp[i] \leftarrow dp[i-k^*] + W(P_{i-k^*+1:i})$ 
9:    $path[i] \leftarrow path[i-k^*] \cup \{k^*\}$ 
10: end for
11:  $\mathcal{R} \leftarrow \emptyset, \quad idx \leftarrow 1$ 
12: for  $k \in path[L]$  do
13:    $\mathcal{R} \leftarrow \mathcal{R} \cup \{\text{MakeRPP}(P_{idx:idx+k-1})\}$ 
14:    $idx \leftarrow idx + k$ 
15: end for
16: return  $\mathcal{R}$ 

```

each note based on its metrical grid position, duration, and syncopation pattern.

3.2 Hierarchical Sequence Representation

To overcome self-attention dilution in long sequences [15], we propose a hierarchical generation architecture that strictly decouples the RPP-level and Note-level.

3.2.1 RPP-Level Representation

We model the global structure as a sequence of Rhythm-Pitch Primitives. The sequence is $\mathcal{V} = \{v_1, \dots, v_T\}$. Each unit is explicitly defined as $v_i = \{b_i, p_i, d_i, r_i, m_i\}$, where b_i , p_i , and d_i denote the measure index, metrical position, and duration respectively. The rhythmic pattern r_i and melodic contour m_i characterize the motif. This abstraction achieves significant temporal downsampling, enabling efficient identification of form development without external priors.

3.2.2 Note-Level Representation

To realize the RPP-level representation as Note-level outputs, we define a note sequence $\mathcal{U} = \{u_1, \dots, u_K\}$ mapping to final MIDI events. Each unit $u_j = \{b_j, p_j, d_j, n_j\}$ shares the aforementioned temporal grid, while n_j explicitly defines the absolute pitch.

3.2.3 Compound Word Embedding Strategy

Extending the Compound Word parallel fusion concept [48], the discrete attributes within v_i and u_j are independently mapped into dense vectors at each time step. These vectors are subsequently aggregated into a unified hidden representation via concatenation and linear projection.

3.3 Model

As shown in Figure 2, our two-stage hierarchical generation framework decouples melody construction into RPP-level structural planning and Note-level instantiation. This

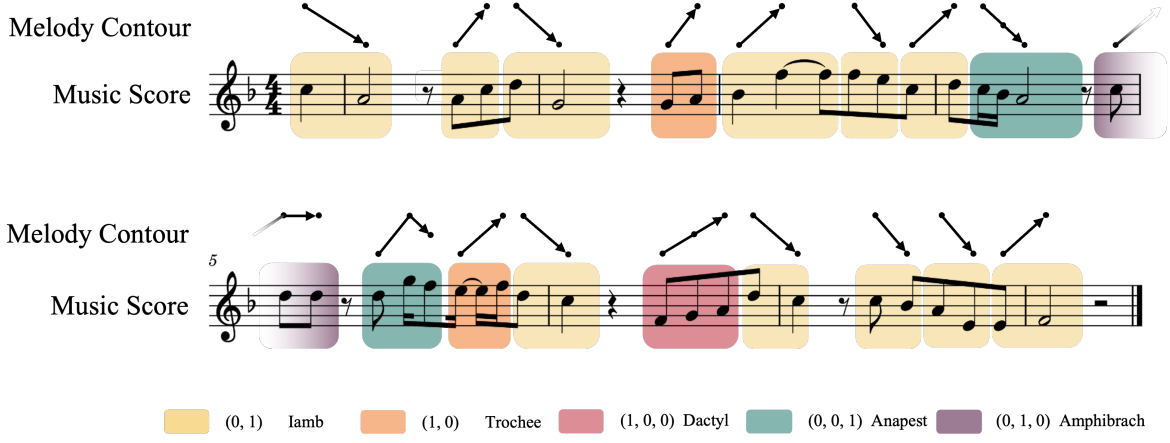


Figure 1. Example of RPP grouping (excerpt from Hey Jude). Colored legends indicate rhythmic types; black arrows denote melodic contour direction.

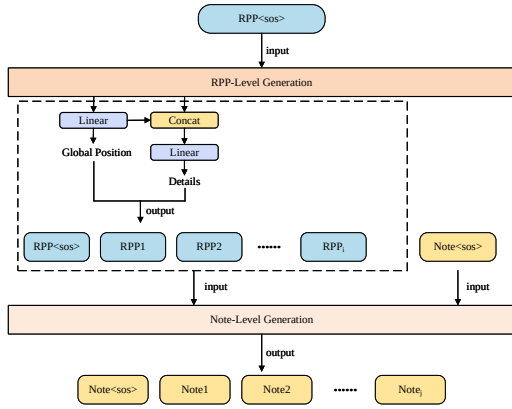


Figure 2. The architecture of the proposed two-stage hierarchical generation framework. Melody construction is strictly decoupled into RPP-level structural planning and Note-level note instantiation. The dashed box details the decoupled serial prediction mechanism deployed at the output end.

multi-resolution temporal representation effectively addresses the semantic disparity between abstract structural development and Note-level symbolic generation.

3.3.1 RPP-Level Generation

The RPP-level structural module aims to generate a long-term framework guiding musical development, featuring a standard Transformer encoder-decoder architecture as its core [34].

While employing compound representation inputs to compress sequence length, we replace traditional parallel prediction with a decoupled serial prediction mechanism at the output, highlighted by the dashed box in Figure 2. Given historical context features $\mathbf{h}_{<t}$ at time step t , the joint probability of primitive v_t is strictly decoupled into the marginal probability of temporal position pos_t and conditional probabilities of detailed attributes, formulated in

Eqn (1):

$$P(v_t|\mathbf{h}_{<t}) = P(pos_t|\mathbf{h}_{<t}) \cdot P(d_t, r_t, m_t|\mathbf{h}_{<t}, pos_t) \quad (1)$$

The model first predicts pos_t solely from $\mathbf{h}_{<t}$. This position representation is subsequently mapped into a temporal prior embedding and concatenated with $\mathbf{h}_{<t}$, enabling the joint prediction of duration, rhythmic pattern, and melodic contour under explicit temporal constraints.

3.3.2 Note-Level Generation

The Note-level module decodes the RPP sequence into concrete note events via a Transformer encoder-decoder [34]. We construct a time-scale expansion mapping that bridges discrete RPP events and the continuous note-level time axis: for each primitive i , the absolute onset is computed as $T_{start} = b_i \times \mathcal{R}_{bar} + p_i$, and its hidden state $\mathbf{h}_i$ is broadcast across all grid steps within $[T_{start}, T_{start} + d_i)$. This allows the decoder to attend to structural memory via cross-attention, ensuring strict alignment between output notes and the RPP-level skeleton.

4. EXPERIMENT

4.1 Dataset

We adopt the MelodyNet dataset [49], comprising over 300,000 MIDI melodies from FreeMIDI, HookTheory, BitMIDI, MuseScore, KernScores, and Kunstderfuge. After removing non-melody tracks and incomplete data, 274,300 melodies remain, split into train/validation/test sets at a ratio of 18 : 1 : 1.

4.2 Compared Models

To evaluate the structural awareness of RPPNet, we select two representative baseline models, human-composed melodies, and two framework variants to validate the specific contribution of the proposed heuristic RPP grouping:

- Museformer [15]: A long-sequence music generation model with fine- and coarse-grained attention.

- MELONS [17]: A graph-sequence architecture that models bar-level chord and melodic contours via line graphs to condition token generation.
- RPPNet-Real: A variant using ground-truth RPP sequences as RPP-level input, representing the theoretical ceiling of Note-level generation.
- RPPNet-Random-Grouped: An ablation baseline replacing heuristic RPP grouping with random segmentation while maintaining architectural consistency and length distribution alignment.

We select Museformer [15] and MELONS [17] as primary baselines, as both exemplify fixed bar-level structural modeling — the exact assumption our work relaxes. MelodyGLM [49] is excluded because its pretrain-finetune paradigm confounds structural boundary design with large-scale pretraining. PhraseLDM [36] is excluded as it requires explicit phrase-level annotations unavailable in MelodyNet. Our ablation baseline RPPNet-Random-Grouped isolates the contribution of perceptual semantics from mere variable-length grouping, yielding a clean controlled comparison: fixed-bar (Museformer, MELONS) vs. perceptually-grouped variable-length (RPPNet) vs. randomly-grouped variable-length (ablation).

4.3 Experiment Setup

4.3.1 Objective Metrics

To validate the effectiveness of RPP grouping, we construct a random-grouping baseline (Random-Grouped) that preserves the length distribution of the RPP training set via random sampling while removing structural semantics.

Following Museformer [15], we adopt Perplexity (PPL) and Structural Error (SE) as objective metrics. This alignment isolates the effect of structural representation choice from metric selection bias. PPL requires a consistent vocabulary across models, so Museformer and MELONS are excluded from the objective comparison. We also note growing evidence questioning the perceptual validity of conventional symbolic music objective metrics [50].

We further verify the RPP-level generator’s learning of underlying distributions by comparing attribute statistics between ground-truth and generated RPP sequences.

4.3.2 Subjective Metrics

Following [49], we conduct a subjective listening test with 15 paid participants (7 with musical background, 8 without) rating generated melodies on a 10-point scale (1–10). All models are trained from scratch, yielding 50 32-bar melodies evaluated across five configurations—RPPNet, RPPNet-Real, RPPNet-Random-Grouped, Museformer, and MELONS—alongside 10 human-composed references from the test set.

Participants were asked to rate each melody on a scale of 1–10 (lowest to highest) along the following four dimensions:

- Coherence: Is the melody fluent, pleasant, and engaging?

- Rhythmicity: Does the melody exhibit a regular metric pattern and appropriate use of rests?
- Structure: Does the melody display clear structural features, such as reasonable repetition and motivic development?
- Overall Impression: A holistic evaluation of the melody.

We select Museformer [15] and MELONS [17] as primary baselines, both representing fixed bar-level structural modeling that our work relaxes. MelodyGLM [49] is excluded as its pretrain-finetune paradigm introduces a confounding variable unrelated to boundary design. PhraseLDM [36] is excluded due to its reliance on explicit phrase annotations unavailable in MelodyNet. The ablation baseline RPPNet-Random-Grouped isolates perceptual semantics from variable-length grouping alone, enabling a controlled comparison across fixed-bar, perceptually-grouped, and randomly-grouped conditions.

5. RESULT

5.1 Statistic Analysis

To quantify RPP-level distributional fidelity, we statistically compare real and generated attribute distributions, yielding a Pearson correlation of 92.19% across 23 categories (Table 2). Dimension-level analysis reveals strong alignment in note group ($r = 0.98$) and melodic contour ($r = 0.90$), while rhythm pattern shows the largest gap ($r = 0.44$), primarily due to over-generation of iambs (0,1: +15.5%) and under-generation of trochees (1,0: −11.4%). These results suggest the model effectively captures note grouping and pitch contour structure, while rhythmic pattern fidelity remains the primary direction for improvement.

5.2 Ablation Study

To validate the effectiveness of heuristic RPP grouping, we construct a random-grouping ablation baseline (Random-Grouped) within the RPPNet framework with aligned length distribution. Objective metrics (Table 3) show that standard RPPNet ($PPL = 2.21$, $SE = 0.0132$) outperforms Random-Grouped (2.36, 0.0175).

Subjective evaluation (Table 1) further reveals that Random-Grouped exhibits significant degradation across all dimensions upon removing heuristic grouping: coherence (5.99, $p < 0.001$), rhythmic quality (6.08, $p < 0.01$), structural quality (6.01, $p < 0.001$), and overall impression (6.09, $p < 0.01$). This demonstrates that the structural semantics captured by heuristic grouping are critical for melodic coherence and motif organization.

Meanwhile, standard RPPNet and RPPNet-Real (ground-truth RPP input) show no significant difference in coherence (6.61 vs. 6.61, $p = 1.00$), rhythmic quality (6.61 vs. 6.59, $p = 0.93$), or overall impression (6.69 vs. 6.61, $p = 0.55$), with structural quality also approaching

Model	Coherence	Rhythmicity	Structure	Overall
Human	$7.29 \pm 0.74^{**}$	$7.36 \pm 0.68^{**}$	$7.45 \pm 0.82^{**}$	$7.47 \pm 0.79^{**}$
RPPNet (Ours)	6.61 ± 0.71	6.61 ± 0.69	6.77 ± 0.73	6.69 ± 0.68
RPPNet-Real	6.61 ± 0.85	6.59 ± 0.83	6.51 ± 0.98	6.61 ± 0.87
RPPNet-Rnd-Grp	$5.99 \pm 0.62^{***}$	$6.08 \pm 0.54^{**}$	$6.01 \pm 0.64^{***}$	$6.09 \pm 0.66^{**}$
Museformer [15]	$5.81 \pm 0.54^{***}$	$5.62 \pm 0.53^{***}$	$5.57 \pm 0.58^{***}$	$5.73 \pm 0.54^{***}$
MELONS [17]	$5.39 \pm 0.69^{***}$	$5.33 \pm 0.66^{***}$	$5.31 \pm 0.76^{***}$	$5.30 \pm 0.77^{***}$

Table 1. Results of the subjective evaluation. Scores are presented as Mean $\pm$ Standard Deviation. Asterisks denote significant differences compared to our proposed RPPNet ($^{**}p < 0.01$, $^{***}p < 0.001$). “Overall” denotes the overall musical impression.

Attribute / Pattern	Real (%)	Generated (%)
<i>Note Group</i>		
One note	12.26	10.61
Two notes	36.25	40.41
Three notes	51.49	48.98
<i>Rhythm Pattern</i>		
(0,)	7.74	8.43
(1,)	4.51	2.19
(0, 1)	14.29	29.83
(1, 0)	21.96	10.57
(0, 0, 1)	14.14	16.31
(0, 1, 0)	17.72	17.18
(1, 0, 0)	19.64	15.49
<i>Melodic Contour</i>		
Descending	15.47	15.15
Ascending	13.18	18.19
Stationary	7.60	7.06
Cont. Descending	7.68	6.83
Cont. Ascending	7.26	6.28
Cont. Stationary	6.32	3.25
Down then Up	8.93	8.91
Up then Down	10.11	8.15
Stay then Up	2.85	4.46
Stay then Down	3.30	3.98
Down then Stay	2.51	3.74
Up then Stay	2.54	3.39
Single Note	12.26	10.61

Table 2. Detailed statistical distributions of note density, rhythm pattern, and melodic contour across real and generated datasets.

non-significance (6.77 vs. 6.51, $p = 0.08$). This suggests that the RPP-level generator of RPPNet possesses strong distribution fitting capabilities, corroborating the findings from statistical experiments.

5.3 Comparison with Previous Models

Table 1 presents the subjective ratings and significance tests against baselines. Human-composed melodies score 7.47 overall as the reference ceiling. RPPNet achieves an overall impression of 6.69, significantly outperforming Museformer (5.73, $p < 0.001$) and Melons (5.30, $p < 0.001$). The advantage is most pronounced in structural quality: RPPNet (6.77) exceeds Museformer (5.57, $p < 0.001$) and MELONS (5.31, $p < 0.001$) by 1.20 and 1.46 points, respectively, demonstrating that explicit hierarchical structural modeling effectively enhances the organization of long-term structure.

Model	PPL $\downarrow$	SE $\downarrow$
RPPNet (Ours)	2.21	0.0132 ± 0.0013
RPPNet-Random-Grouped	2.36	0.0175 ± 0.0008

Table 3. Results of the objective evaluation in the ablation study. SE denotes the structural entropy of Rhythmic Pattern Primitives (RPPs). The best results are highlighted in bold.

6. DISCUSSION

This work integrates music-psychological principles with music theory to propose RPPNet, a perception-driven structural representation for melody generation that transcends fixed bar-level boundaries through hierarchical decoupled generation. Experiments confirm that this framework significantly outperforms existing baselines in long-term structural coherence and overall musicality.

These results demonstrate that perception-driven grouping significantly outperforms fixed bar-level representations. Ablation experiments reveal that performance gains stem from the proper encoding of perceptual grouping principles—merely breaking bar isometry without respecting perceptual rules yields no benefit.

Although RPPNet significantly outperforms existing baselines, a quality gap remains compared to human compositions (overall 7.47 vs. 6.69), likely attributable to the distributional mismatch in rhythmic patterns (Pearson $r = 0.44$). Closing this gap is the central challenge going forward.

Future work will proceed along three directions: (1) hybrid data-driven and rule-driven structural strategies; (2) introducing micro-macro feedback mechanisms where Note-level generation revises RPP-level structure; and (3) extending this representation to polyphonic music and accompaniment generation, evaluating its capability under complex musical textures.

7. REFERENCES

- [1] S. Ji, X. Yang, and J. Luo, “A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges,” *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–39, 2023.
- [2] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synaev, Y. Adi, and A. Défossez, “Simple and control-

- lable music generation,” *Advances in neural information processing systems*, vol. 36, pp. 47 704–47 720, 2023.
- [3] J. Liang, H. Zhang, H. Liu, Y. Cao, Q. Kong, X. Liu, W. Wang, M. D. Plumbley, H. Phan, and E. Benetos, “Wavcraft: Audio editing and generation with large language models,” in *ICLR 2024 Workshop on LLM Agents*, 2024. [Online]. Available: <https://openreview.net/forum?id=xJw7x2ZBex>
 - [4] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, “Audioldm: Text-to-audio generation with latent diffusion models,” *arXiv preprint arXiv:2301.12503*, 2023.
 - [5] J.-P. Briot and F. Pachet, “Deep learning for music generation: Challenges and directions,” *Neural Computing and Applications*, vol. 32, no. 4, pp. 981–993, 2020.
 - [6] H. C.-Z. Anna, V. Ashish, U. Jakob, S. Ian, H. Curtis, S. Noam, D. Monica, E. Douglas *et al.*, “Music transformer: Generating music with long-term structure,” *arXiv preprint*, 2018.
 - [7] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko *et al.*, “Highly accurate protein structure prediction with AlphaFold,” *nature*, vol. 596, no. 7873, pp. 583–589, 2021.
 - [8] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, “Neural speech synthesis with transformer network,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6706–6713.
 - [9] P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, and T. Laino, “Extraction of organic chemistry grammar from unsupervised learning of chemical reactions,” *Science Advances*, vol. 7, no. 15, p. eabe4166, 2021.
 - [10] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, “Medical transformer: Gated axial-attention for medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 36–46.
 - [11] F. Carnovalini and A. Rodà, “Computational creativity and music generation systems: An introduction to the state of the art,” *Frontiers in Artificial Intelligence*, vol. 3, p. 14, 2020.
 - [12] Y. Ma, A. Øland, A. Ragni, B. M. Del Sette, C. Saitis, C. Donahue, C. Lin, C. Plachouras, E. Benetos, E. Shatri *et al.*, “Foundation models for music: A survey,” *arXiv preprint arXiv:2408.14340*, 2024.
 - [13] Z. Yin, F. Reuben, S. Stepney, and T. Collins, “Deep learning’s shallow gains: A comparative evaluation of algorithms for automatic music generation,” *Machine Learning*, vol. 112, no. 5, pp. 1785–1822, 2023.
 - [14] R. Bader, *Springer Handbook of Systematic Musicology*. Springer, 2018.
 - [15] B. Yu, P. Lu, R. Wang, W. Hu, X. Tan, W. Ye, S. Zhang, T. Qin, and T.-Y. Liu, “Museformer: Transformer with fine-and coarse-grained attention for music generation,” *Advances in neural information processing systems*, vol. 35, pp. 1376–1388, 2022.
 - [16] Y.-J. Shih, S.-L. Wu, F. Zalkow, M. Müller, and Y.-H. Yang, “Theme transformer: Symbolic music generation with theme-conditioned transformer,” *IEEE Transactions on Multimedia*, vol. 25, pp. 3495–3508, 2022.
 - [17] Y. Zou, P. Zou, Y. Zhao, K. Zhang, R. Zhang, and X. Wang, “Melons: Generating melody with long-term structure using transformers and structure graph,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 191–195.
 - [18] J. Wu, X. Liu, X. Hu, and J. Zhu, “PopMNet: Generating structured pop music melodies using neural networks,” *Artificial Intelligence*, vol. 286, p. 103303, 2020.
 - [19] G. W. Cooper, G. Cooper, and L. B. Meyer, *The Rhythmic Structure of Music*. University of Chicago press, 1963.
 - [20] I. Deliege, “Grouping conditions in listening to music: An approach to Lerdaahl & Jackendoff’s grouping preference rules,” *Music perception*, vol. 4, no. 4, pp. 325–359, 1987.
 - [21] A. Lamont and N. Dikken, “Motivic structure and the perception of similarity,” *Music Perception*, vol. 18, no. 3, pp. 245–274, 2001.
 - [22] Y. Ren, J. He, X. Tan, T. Qin, Z. Zhao, and T.-Y. Liu, “Popmag: Pop music accompaniment generation,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1198–1206.
 - [23] J. Edworthy, “Interval and contour in melody processing,” *Music Perception*, vol. 2, no. 3, pp. 375–388, 1985.
 - [24] —, “Pitch and contour in music processing,” *Psychomusicology: A Journal of Research in Music Cognition*, vol. 2, no. 1, p. 44, 1982.
 - [25] F. P. Brooks, A. L. Hopkins, P. G. Neumann, and W. V. Wright, “An experiment in musical composition,” *IRE Transactions on Electronic Computers*, no. 3, pp. 175–182, 1957.
 - [26] P. Mavromatis, “A hidden Markov model of melody production in Greek church chant,” *Computing in musicology: a directory of research*, no. 14, pp. 93–112, 2006.

- [27] —, “HMM Analysis of Musical Structure: Identification of Latent Variables Through Topology-Sensitive Model Selection,” in *International Conference on Mathematics and Computation in Music*. Springer, 2009, pp. 205–217.
- [28] K. P. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*. University of California, Berkeley, 2002.
- [29] H. Chu, R. Urtasun, and S. Fidler, “Song From PI: A Musically Plausible Network for Pop Music Generation,” in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.
- [30] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. C. Courville, and Y. Bengio, “A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4–9, 2017, San Francisco, California, USA*, S. Singh and S. Markovitch, Eds. AAAI Press, 2017, pp. 3295–3301.
- [31] L.-C. Yang, S.-Y. Chou, and Y.-H. Yang, “MidiNet: A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation,” in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23–27, 2017*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 324–331.
- [32] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. Le, and R. Salakhutdinov, “Transformer-xl: Attentive language models beyond a fixed-length context,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2978–2988.
- [33] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, “Transformers are rnns: Fast autoregressive transformers with linear attention,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 5156–5165.
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [35] N. Fradet, J.-P. Briot, F. Chhel, A. E. F. Seghrouchni, and N. Gutowski, “Meditok: A python package for midi file tokenization,” *arXiv preprint arXiv:2310.17202*, 2023.
- [36] L. Ou and Y. Wang, “PhraseVAE and PhraseLDM: Latent Diffusion for Full-Song Multitrack Symbolic Music Generation,” *CoRR*, vol. abs/2512.11348, 2025.
- [37] N. Fradet, N. Gutowski, F. Chhel, and J.-P. Briot, “Byte pair encoding for symbolic music,” in *Proceedings of the 2023 conference on empirical methods in natural language processing*, 2023, pp. 2001–2020.
- [38] S.-L. Wu and Y.-H. Yang, “The Jazz Transformer on the Front Line: Exploring the Shortcomings of AI-composed Music through Quantitative Measures,” in *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11–16, 2020*, J. Cumming, J. H. Lee, B. McFee, M. Schedl, J. Devaney, C. McKay, E. Zangerle, and T. de Reuse, Eds., 2020, pp. 142–149.
- [39] M. Zeng, X. Tan, R. Wang, Z. Ju, T. Qin, and T.-Y. Liu, “Musicbert: Symbolic music understanding with large-scale pre-training,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 791–800.
- [40] X. Zhang, J. Zhang, Y. Qiu, L. Wang, and J. Zhou, “Structure-enhanced pop music generation via harmony-aware learning,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1204–1213.
- [41] G. Wu, S. Liu, and X. Fan, “The power of fragmentation: A hierarchical transformer model for structural segmentation in symbolic music generation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1409–1420, 2023.
- [42] S. Dai, Z. Jin, C. Gomes, and R. B. Dannenberg, “Controllable deep melody generation via hierarchical music structure representation,” in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7–12, 2021*, J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. van Kranenburg, and A. Srinivasamurthy, Eds., 2021, pp. 143–150.
- [43] F. Jin, P. Li, and X. Wu, “A theme music generation model based on hybrid variational autoencoders and conditional generative adversarial networks,” *Engineering Applications of Artificial Intelligence*, vol. 161, p. 112131, 2025.
- [44] A. Forte and S. E. Gilbert, “Introduction to Schenkerian analysis,” (*No Title*), 1982.
- [45] S. Koelsch, M. Rohrmeier, R. Torrecuso, and S. Jentschke, “Processing of hierarchical syntactic structure in music,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 38, pp. 15 443–15 448, 2013.
- [46] F. Lerdahl and R. S. Jackendoff, *A Generative Theory of Tonal Music, Reissue, with a New Preface*. MIT press, 1996.
- [47] M. Rohrmeier and P. Rebuschat, “Implicit learning and acquisition of music,” *Topics in cognitive science*, vol. 4, no. 4, pp. 525–553, 2012.
- [48] W.-Y. Hsiao, J.-Y. Liu, Y.-C. Yeh, and Y.-H. Yang, “Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs,”

in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 178–186.

- [49] X. Wu, Z. Huang, K. Zhang, J. Yu, X. Tan, T. Zhang, Z. Wang, and L. Sun, “MelodyGLM: Multi-task Pre-training for Symbolic Melody Generation,” *CoRR*, vol. abs/2309.10738, 2023.
- [50] A. Lerch, C. Arthur, N. Bryan-Kinns, C. Ford, Q. Sun, and A. Vinay, “Survey on the evaluation of generative models in music,” *ACM Comput. Surv.*, vol. 58, no. 4, Oct. 2025. [Online]. Available: <https://doi.org/10.1145/3769106>