Representation Matters: An Evaluation of MEL, CQT, VQT, and MCQT for Beat Tracking

Joaquim B. Cavalcante[24] , Julio Hsu[14] , Carlos Batista[2] , Telmo M. Filho[3] , Thaís Gaudêncio[2] , Yuri Malheiros[2] ,

[1]Centro de Ciências e Tecnologia - UFCG
Rua Aprígio Veloso, 882 - Campina Grande - PB

[2]Centro de Informática - UFPB
Cidade Universitária - João Pessoa - PB

[3]University of Bristol
Bristol, United Kingdom

[4]Music AI
Salt Lake City, UT, USA

julio.hsu@ccc.ufcg.edu.br, joaquim.breno@academico.ufpb.br, yuri@ci.ufpb.br

Abstract. *Neural beat-tracking models predominantly rely on Mel spectrograms. These representations often fail to capture harmonic nuances in complex genres (e.g., jazz, classical), leading to rhythmically ambiguous signal patterns and suboptimal performance. We challenge the architectural-centric paradigm in music analysis by (1) evaluating alternative time-frequency representations, and (2) proposing a dual-input architecture to synergize complementary spectrogram properties. Through controlled experiments, we compare Mel, Constant-Q Transform (CQT), and Variable-Q Transform (VQT) inputs. We then introduce MCQT: a dual-branch model that fuses Mel and CQT, validated using 8-fold cross-validation and ensemble inference. Replacing Mel with CQT reduced subharmonic errors by 22% in harmony-rich contexts. MCQT further outperformed single-input baselines, achieving state-of-the-art performance in annotation coverage (0.912 vs. Mel's 0.910) and reducing subharmonic errors by 26% compared to Mel. It excelled in complex genres (e.g., 6.4% higher downbeat F1-score in classical). Strategic representation engineering, not just architectural complexity, is pivotal for robust beat tracking. MCQT's fusion of complementary spectrograms establishes a new paradigm for MIR tasks, particularly in harmonically complex music.*

1. Introduction

How well a model performs often depends not only on its architecture but also critically on the nature of its input representation [1]. In many audio processing tasks, especially those involving musical structure, the choice of spectrogram representation can drastically influence outcomes [2]. Different representations offer distinct advantages: the Mel spectrogram, with its roots in speech processing, captures perceptual loudness cues well and is computationally efficient, making it suitable for voice-based applications. The Constant-Q Transform (CQT), by contrast, aligns more closely with musical pitch perception due to its logarithmic frequency scaling, making it better suited for generic music tasks. [3]. The Variable-Q Transform (VQT), offering even more flexible resolution, shows potential for specialized musical material, though it may demand more expressive models to realize its benefits [4].

In the field of computational musicology in Music Information Retrieval (MIR), neural beat tracking has become a cornerstone, enabling applications ranging from automatic transcription to interactive music systems. Recent advancements have predominantly leveraged sequence-to-sequence architectures, often employing Mel spectrograms as input representations due to their success in speech processing tasks [5, 6]. However, this approach inherits the limitations of the Mel design: the Mel spectrogram's linear frequency resolution and fixed time-frequency granularity may not adequately capture the intricate harmonic and rhythmic structures inherent in complex musical genres such as classical piano and jazz [7].

Motivated by these representational limitations, prior research has explored alternative spectrogram designs. Gu et al [8] proposed a Multi-Scale Sub-Band Constant-Q Transform (MS-SB-CQT) discriminator, demonstrating that CQT-based frequency scaling yields more accurate harmonic modeling in vocoders than traditional Short-Time Fourier Transform (STFT) approaches. Similarly, Guo et al [9] introduced "Mel-Refine", a plug-and-play enhancement method that improves the fidelity of Mel spectrograms in generative tasks by rebalancing component contributions—revealing that even minor refinements in representation can have major perceptual effects. On the other hand, Silva [10] showed that the selection of psychoacoustic descriptors plays a central role in the analytical power of computational musicology frameworks, emphasizing the importance of representations that align with human auditory perception and reinforcing the idea that the success of a model often hinges on the relevance of its input features to the task at hand. Their work underscores the necessity of selecting appropriate features to effectively capture the nuances of musical structures.

These results make it evident that the choice of input representation significantly influences a model's capacity to perceive and process musical structure. As the spectral and temporal complexity of musical content increases, particularly in harmony-rich genres such as classical piano, the model's ability to accurately distinguish rhythmic cues and structural elements becomes

increasingly dependent on the fidelity of the representation. Therefore, time-frequency representations such as the Constant-Q Transform offer a more appropriate foundation for modeling these complex auditory phenomena. To investigate this hypothesis, we retrain the BeatThis neural beat-tracking model [11] from scratch using 8-fold cross-validation across three spectrogram inputs: Mel, CQT, and VQT. Later on, we introduce a novel dual-path architecture following established multimodal fusion approaches, where separated Mel and CQT frontends extract complementary features that are then combined through early fusion. A Mixture of Experts module then processes these concatenated features, allowing the model to learn when to rely on each representation type. This multimodal approach allows us to evaluate not only the individual effectiveness of each representation but also the potential benefits of combining complementary time-frequency representations within a unified neural model for improved beat tracking performance.

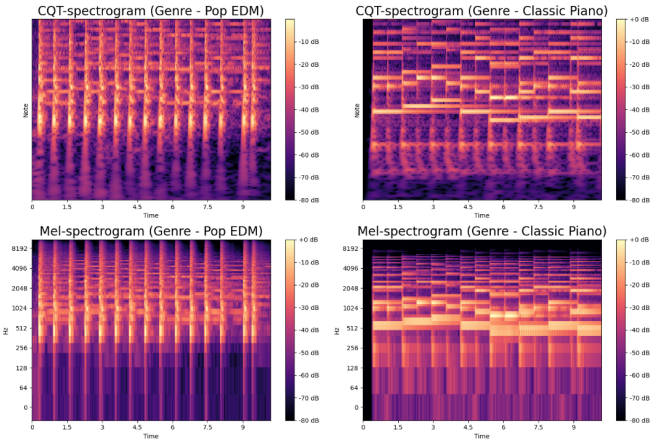

Figure 1: Spectrogram comparison across Mel and CQT representations (rows-axis) and musical genres Pop/EDM and Classical Piano (columns-axis).

2. Literature Review

This section provides a comprehensive overview of limitations with relevant research around spectrogram representations and beat tracking in music signal processing.

2.1. Spectrogram Representations in Music Signal Processing

Spectrogram representations are fundamental in audio signal processing, serving as critical inputs for various MIR tasks [2]. Among these, the Mel spectrogram and the CQT are prominent due to their distinct characteristics and applications.

The Mel spectrogram, derived from the Short-Time Fourier Transform (STFT), applies a filter bank spaced according to the Mel scale, which approximates the human ear's perception of pitch [5, 12]. This representation emphasizes lower frequencies, aligning well with speech processing tasks. Its fixed time-frequency resolution and compact representation make it computationally efficient and suitable for applications where percussive and rhythmic elements dominate.

In contrast, the CQT provides a logarithmic frequency resolution, with frequency bins spaced geometrically to align with musical intervals, such as semitones. This structure yields higher frequency resolution at lower frequencies and improved temporal resolution at higher frequencies, closely reflecting the nature of musical signals. Due to its adaptability, the CQT is especially effective in analysing harmonic content and pitch-related features in music [3]. As illustrated in Figure 1,the CQT provides a more refined representation than the Mel spectrogram, particularly in complex musical pieces such as classical piano.

2.2. Beat Tracking and Representation Challenges

Beat tracking—the task of identifying temporal beat positions in music—is a core challenge in MIR, with applications in synchronization, segmentation, and genre classification [13]. However, current systems often suffer from representational limitations that hinder performance across diverse musical styles.

Most neural beat trackers rely on Mel spectrograms due to their efficiency and effectiveness in rhythm-focused genres like pop and EDM. Yet, this representation struggles with harmonic-rich music—such as classical piano or jazz—because its linear frequency resolution and perceptually motivated scaling fail to capture fine harmonic detail [11, 14]. In such genres, Mel-based inputs often produce ambiguous signals, reducing downbeat detection accuracy.

The state-of-the-art BeatThis system [11], while achieving high F1 scores through a convolutional-transformer design and removing Dynamic Bayesian Network postprocessing, still falters in harmony-intensive contexts. Its poor performance on continuity metrics highlights the limitations of relying solely on Mel spectrograms.

3. Methodology

Our approach systematically isolates representational impact by: (1) replacing Mel spectrograms with CQT and VQT representations in the original BeatThis architecture (2) introducing a novel dual-input MCQT model that synergizes Mel's rhythmic precision with CQT's harmonic richness. Our complete implementation, including training scripts and model architectures, is publicly available [15]. This representation-centric methodology challenges the field's architectural focus, demonstrating that strategic input engineering can yield substantial performance gains without structural model complexity. The dual-input strategy directly addresses the fundamental trade-off between temporal and spectral resolution in time-frequency representations. By leveraging complementary spectral properties through expert-based fusion, we transform representational limitations into architectural advantages—establishing a new paradigm for robust beat tracking across diverse musical contexts.

3.1. Datasets for Training

While BeatThis focused on a set of 4,556 audio files, our experiments span 9 diverse datasets (see Table 1), collectively comprising 4,645 audio files annotated with both beat and downbeat information, amounting to a total of 145.55 hours of music.

Dataset	Tracks	Minutes	Time (h:mm)
Frank	1,984	5,709.66	95:10
Groovemidi [16]	1150	604.50	10:05
Simac [17]	595	193.18	3:13
HJDB [18]	235	197.81	3:18
Hainsworth [19]	222	197.84	3:18
SMC [20]	217	142.35	2:22
Jaah [21]	113	405.41	6:45
TapCorrect [22]	81	410.49	6:50
Filosax [23]	48	294.22	04:55
Total	**4,645**	**8,755.46**	**145:55**

Table 1: All of Dataset Statistics for Beat and Downbeat Tracking.

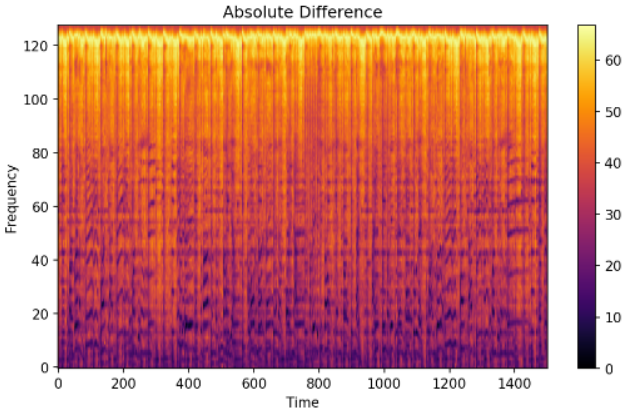

Figure 2: Highlighting variations in magma-colored from two spectrograms, where lighter regions indicate greater variation.

Notably, the Frank dataset—comprising 1,984 tracks and totaling approximately 95 hours of audio—represents a large-scale, balanced corpus assembled from a diverse collection of sources, including GuitarSet [24], Beatles [25], RWC [26], Ballroom [27], ASAP [28], Harmonix [29], and Candombe [30].

Consistent with the BeatThis framework [11], we apply the same data augmentation techniques to enhance generalization. These include pitch shifting (±6 semitones), tempo scaling (±20%), and a temporal masking strategy, where 0.5–2 second segments are randomly reordered.

3.2. Model and Training

To ensure fair and rigorous evaluation of our proposed methods, we retrained the original BeatThis model from scratch using 8-fold cross-validation with datasets mentioned at session 3.1, augmentation techniques, and preprocessing pipeline as described by the original authors. We applied this retraining to all three input representations: Mel, CQT, and VQT spectrograms respectively.

As detailed in Section 3.1 (Dataset Training), we used a slightly larger dataset than the original authors. Additionally, as shown in Figure 2, generating Mel spectrograms from the same audio input at different times introduces non-deterministic variations. The absolute differences between spectrograms from successive runs, although minor, reveal systematic inconsistencies that affect reproducibility and lead to variability in performance scores—ultimately undermining fair comparison. These variations introduce enough noise to compromise reliable evaluation across experimental conditions. Therefore, it is essential to establish our own controlled preprocessing pipeline, rather than relying on potentially inconsistent portions of pretrained aggregated datasets and model comparisons.

The full experimental pipeline were conducted in two main steps of data setup:

Phase One – We trained the model on a reduced and deliberately imbalanced dataset, by not including the GrooveMIDI (percussive-heavy) and Filosax (harmonic-heavy) subsets, in order to test generalization under constrained less data conditions. This phase compared model performance across Mel, CQT, and VQT inputs using the original BeatThis architecture. Due to VQT's consistently poor results (see Table 2), it was excluded from further experiments. We hypothesize that VQT's greater representational variability demands a more complex or specialized architecture, and expect future research in this direction.

Phase Two – We trained from scratch on the full dataset (more data than phase one) shown at Table 1, while in this phase still comparing Mel, CQT, but in place of VQT we propose a newly dual-input architecture, as shown in Figure 3.

Our proposed dual-path architecture draws inspiration from multimodal learning research demonstrating how different sensory modalities—visual, auditory, and tactile—can be processed through specialized pathways before integration for enhanced perception and decision-making [31]. Reflecting this paradigm, our model employs parallel Mel and CQT pipelines that act as specialized feature extractors, integrates their complementary representations via a Mixture of Experts (MoE) module [32], feeding the unified result into the core BeatThis architecture. The MoE framework uses four experts and two gating networks (K=2) to dynamically weight Mel or CQT features based on musical context. By explicitly modeling the interplay of diverse time-frequency representations before unified decision-making, the architecture enables a rigorous evaluation of whether combining multiple spectral views improves beat tracking accuracy and robustness across different musical genres and acoustic conditions.

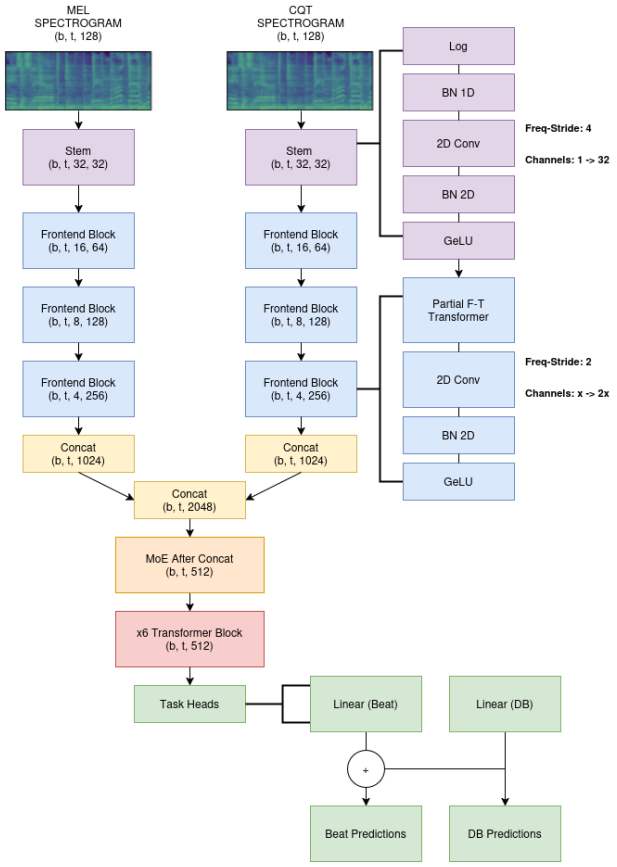

Figure 3: MCQT dual-input architecture.

3.3. Evaluation

We evaluate beat tracking systems on the unseen GTZAN dataset [33], using key metrics around Beat (B) and Down-Beat (DB). This approach captures both fine-grained temporal accuracy and broader metrical-hierarchical structural cues, enabling statistically robust comparisons and ensuring reliable performance across diverse musical corpora.

3.3.1. Metrics Explanation

To rigorously assess beat tracking performance across various rhythmic complexities and musical contexts, we detail the three primary classes of metrics used in our study—F-Measure (F-M), Continuity Metrics (CMLt/AMLt) [34], and Annotation Coverage Ratio (ACR) [35]—highlighting their definitions, underlying assumptions, and how they reflect different dimensions of beat tracking accuracy.

F-M evaluates beat tracking by considering precision and recall within a fixed time window (default 0.07 seconds). An estimated beat is correct if it falls within this window of a reference beat. The F-measure is: F-measure $= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ where Precision is the ratio of correctly estimated beats to total estimated beats, and Recall is the ratio of correctly estimated beats to total reference beats.

Continuity Metrics evaluate how well estimated beats maintain correct tempo and phase. We focus on two measures: CMLt (Correct Metric Level total accuracy) measures beats at the exact metrical level, while AMLt (Any Metric Level total accuracy) allows metrical variations like double or half tempo [36]. Both metrics require beats to satisfy two conditions:

1 - Phase condition Distance between estimated beat $\hat{b}j$ and reference beat b_i relative to local inter-beat interval must be less than threshold τ_{phase} (typically 0.175): $\frac{|b_i - \hat{b}j|}{b_i - bi-1} < \tau_{\text{phase}}$.

2 - Period condition Ratio between estimated and reference inter-beat intervals must be within threshold τ_{period} (typically 0.175) of 1.0: $\left|1 - \frac{\hat{b}j - \hat{b}j-1}{b_i - b_{i-1}}\right| < \tau_{\text{period}}$ Total accuracy is the proportion of correctly identified beats relative to total estimated beats.

ACR evaluates beat tracking with greater rhythmic flexibility, recognizing that music often supports multiple valid metrical interpretations. While ACR includes standard on-beat and off-beat evaluations, its core strength lies in analyzing subharmonic and harmonic tempo structures. Subharmonic ACR tests whether an algorithm can track slower metrical levels by subsampling reference beats—such as every second (half tempo), third (third tempo), or fourth beat (quarter tempo). Harmonic ACR, in contrast, evaluates faster interpretations by interpolating additional beats between references—doubling, tripling, or quadrupling the tempo. To maintain temporal accuracy at higher tempi, ACR employs an adaptive tolerance window that scales with the inter-beat interval, rather than relying on a fixed margin.

3.3.2. Scoring Strategy

To rigorously assess the performance and robustness of our beat tracking framework, we adopt a multi-tiered evaluation strategy that deliberately deviates from the conventional cross-validation protocols commonly employed in MIR. Instead, we replicate the training and evaluation configuration proposed by the authors of BeatThis, with the specific goal of addressing the harmonic ambiguity problem highlighted in their work. This alignment enables a direct and meaningful comparison, allowing us to demonstrate that our proposed model effectively mitigates the limitations previously identified.

Our scoring methodology is structured around two principal evaluation scopes, designed to assess both generalization capacity and musical plausibility across diverse rhythmic conditions and genres.

1 – Academic Evaluation Scope

Following the BeatThis cross-validation strategy, we evaluate our trained models on the GTZAN, a widely-used, genre-balanced benchmark in MIR. Crucially, GTZAN is entirely disjoint from our training corpus, serving as a strict out-of-distribution testbed to evaluate generalization to unseen musical content. This evaluation includes:

- **Standard Evaluation (F-M, CMLt, AMLt):** We report the mean beat and downbeat accuracies across all GTZAN tracks using the core met-

rics defined by the BeatThis protocol. These metrics assess the model's precision, recall, and continuity in maintaining consistent tempo and phase alignment, providing a comprehensive measure of accuracy under fixed metrical assumptions.

- **Extended Evaluation (ACR):** To account for the inherent rhythmic ambiguity in musical perception, we incorporate Annotation Coverage Ratio (ACR) as a complementary metric. ACR evaluates the model's ability to align with multiple musically plausible metrical interpretations, offering a more flexible and musicologically grounded perspective on beat tracking performance.

2 – Production Evaluation Scope

Ensemble-Based Aggregation: To emulate real-world deployment conditions and enhance prediction stability, we implement an ensemble strategy based on genre-level performance profiling over GTZAN's genre labels. This involves aggregating predictions from all eight cross-validated models using temporal majority voting—retaining a beat event in the final output only if it is supported by the majority of models. This ensemble fusion approach offers several practical advantages, including improved robustness to outlier folds, reduced variance, and increased readiness for production use.

By integrating fold-wise statistical validation, out-of-distribution generalization testing, and ensemble-based prediction aggregation, our scoring strategy establishes a comprehensive and technically rigorous framework for evaluating beat tracking systems. This ensures that the proposed model is not only effective under controlled experimental conditions but also robust, adaptable, and suitable for deployment across diverse real-world musical scenarios.

4. Comparative Study

This section offers a comparative analysis of our methodologies, building on previously outlined evaluation strategies. While Phase One focused on academic evaluation, Phase Two significantly broadens our research through two new, substantial studies. This expansion is due to a larger dataset and a novel model architecture, allowing us to pursue both academic and production-oriented goals.

4.1. Phase One

4.1.1. Academic Evaluation Scope

In this initial phase, we used a dataset characterized by limited percussive and harmonic complexity, by excluding genres such as percussive of GrooveMIDI and harmonic of Filosax. This setting served as a baseline for evaluating the fundamental capabilities of different time-frequency representations.

As shown in Table 2, the Mel spectrogram representation achieved the highest average performance across all evaluated metrics—beat, downbeat, and their mean. Notably, however, the CQT representation outperformed Mel in downbeat prediction in two metrics (CMLt (DB)/AMLt (DB)). This suggests CQT's potential advantage in modeling harmonic structures, which are critical for capturing global rhythmic anchors such as downbeats. This preliminary evidence highlights CQT's utility in encoding long-term harmonic dependencies, a quality less prominent in the Mel representation.

Conversely, the VQT representation underperformed across all metrics. This likely reflects a mismatch between the VQT's highly detailed harmonic encoding and the capacity of the model architecture used, which may not be expressive enough to exploit such complex input features. Although VQT was not pursued further in this study, its potential remains open for future exploration with more advanced models.

Metric	MEL	CQT	VQT
F-M (B)	$\mathbf{87.8 \pm 0.2}$	86.9 ± 0.3	42.3 ± 2.0
F-M (DB)	$\mathbf{75.7 \pm 0.5}$	75.6 ± 0.4	30.2 ± 1.7
CMLt (B)	$\mathbf{77.2 \pm 0.7}$	76.3 ± 0.9	38.9 ± 2.6
CMLt (DB)	62.9 ± 1.1	$\mathbf{63.6 \pm 0.8}$	50.2 ± 1.6
AMLt (B)	$\mathbf{87.4 \pm 0.5}$	86.4 ± 0.5	48.3 ± 2.5
AMLt (DB)	75.0 ± 1.0	$\mathbf{75.8 \pm 1.0}$	59.8 ± 1.9
Overall Mean	**77.7**	77.4	44.9

Table 2: Academic Evaluation Scope (Phase One) – Mean and Standard Deviation of Model Performance.

These findings prompted deeper investigation into the harmonic modeling advantages offered by CQT. Motivated by its promising downbeat detection in low-harmonic contexts, we extended our analysis in Phase Two to include more harmonically rich datasets—such as Filosax jazz—in order to evaluate whether the benefits of CQT would become more pronounced, particularly when fused with Mel in our proposed MCQT dual-input architecture.

4.2. Phase Two

4.2.1. Academic Evaluation Scope

This second phase extends the initial analysis by incorporating musically richer datasets characterized by corporas of high percussive density GrooveMIDI and elevated harmonic complexity Filosax. These datasets serve as a more demanding evaluation, enabling a rigorous evaluation of the scalability and robustness of time-frequency representations under more complex musical structures.

Building upon the insights gained in Phase One, this phase shifts focus to evaluating the dual-branch MCQT architecture, which simultaneously processes Mel-spectrogram and CQT representations. As presented in Table 3, MCQT achieves the highest mean performance across all evaluation metrics—beat accuracy, downbeat accuracy, and their composite mean—consistently outperforming single-representation baselines across all cross-validation folds. This highlights the model's improved

generalization capability and adaptability when exposed to increased musical complexity.

While the single-input models (Mel-only and CQT-only) occasionally demonstrated superior performance in isolated scenarios—e.g., CQT excelling in downbeat prediction. Still, MCQT exhibited consistent superiority in average performance, supporting our hypothesis that combining Mel's superior temporal resolution with CQT's harmonic sensitivity yields a more expressive and balanced representation. This fusion enhances the model's ability to capture both rhythmic transients and harmonic modulations, echoing principles found in multimodal architecture specialized pathways.

Metric	MEL	CQT	MCQT
F-M (B)	87.5 ± 1.8	87.6 ± 1.6	$\mathbf{88.1 \pm 1.8}$
F-M (DB)	76.0 ± 1.9	76.3 ± 3.1	$\mathbf{76.7 \pm 3.1}$
CMLt (B)	76.5 ± 3.0	77.7 ± 3.0	$\mathbf{78.0 \pm 2.5}$
CMLt (DB)	63.1 ± 2.5	$\mathbf{65.4 \pm 3.0}$	63.6 ± 4.5
AMLt (B)	87.7 ± 1.8	87.5 ± 2.0	$\mathbf{87.7 \pm 1.5}$
AMLt (DB)	75.4 ± 2.1	$\mathbf{77.3 \pm 2.9}$	76.6 ± 3.7
Overall Mean	77.70	**78.63**	78.43

Table 3: Academic Evaluation Scope (Phase Two) – Mean and Standard Deviation of Model Performance.

These results reinforce our initial assumptions from Phase One and empirically validate the effectiveness of the CQT model in handling musically intricate contexts, particularly those with pronounced harmonic depth and non-trivial rhythmic patterns.

In parallel, we extend our evaluation with the ACR framework introduced in [35]. As shown in Table 4, the MCQT model achieves the highest overall annotation coverage (0.912) and on-beat accuracy (0.783), outperforming both Mel (0.910, 0.769) and CQT (0.905, 0.779). These results underscore the effectiveness of the dual-input architecture in integrating complementary feature domains for more robust beat tracking.

Critically, while all models maintain low offbeat error rates ($\sim$0.019), the Mel-only model exhibits the highest subharmonic error rate (0.066), followed by CQT (0.059), with MCQT demonstrating the lowest (0.054). This is an important finding, as subharmonic errors—often resulting in tracking at fractional tempos—are perceptually more disruptive than harmonic or phase-shifted deviations [35]. The observed reduction in subharmonic errors in CQT-based models can be attributed to the logarithmic frequency resolution of the Constant-Q Transform, which enhances the preservation of harmonic structure and tempo stability in rhythmically ambiguous passages.

The MCQT model further capitalizes on this advantage by incorporating Mel-based timing precision, ultimately yielding a representation that is both harmonically coherent and temporally accurate. These findings demonstrate that the integration of Mel and CQT representations in the MCQT model not only improves average beat tracking accuracy but also substantially mitigates critical error types, thereby reinforcing its suitability for deployment in rhythmically and harmonically complex musical environments.

4.2.2. Production Evaluation Scope

We conduct a production-level evaluation using ensemble inference on a held-out test set without training data access, simulating real-world deployment conditions. The evaluation operates at both macro and micro levels.

Macro-Level Analysis: As detailed in Table 5, our results demonstrate that MCQT achieves competitive or even superior performance, specifically resulting in a dominance on classical music genres across all metrics. This enhancement is due to the incorporation of CQT into the Mel pipeline, which effectively activates features that single-input models typically miss. MCQT consistently outperforms both standalone Mel and CQT models, validating our central thesis around production environment evaluation: that strategic representation choice and fusion are critical for robust musical structure prediction.

Micro-Level Analysis: And at a micro perspective, a single-genre evaluation reveals that the MCQT ensemble demonstrates a stronger alignment with CQT than with Mel representations. This indicates that the CQT contributes more decisively to the ensemble's predictions than the Mel spectrogram, suggesting that CQT better captures the underlying geometric structure of music. This finding supports the hypothesis that while Mel spectrograms are highly effective for speech tasks, CQT offers superior resolution for musical feature representation.

5. Discussions

Our experiments demonstrate that carefully chosen spectral representations can significantly enhance both beat and downbeat tracking accuracy. In Phase One, Mel spectrograms excelled in overall beat detection on rhythmically complex but harmonically sparse datasets. Then, with the introduction of rhythmically and harmonically dense datasets, CQT outperformed Mel, underscoring its superior harmonic resolution. Finally, by fusing Mel and CQT in our MCQT architecture, we achieved better performance on harmonically/rhythmically complex passages compared to the BeatThis baseline, suggesting that the combined representation helps address some of the challenges posed by rich chordal progressions in the original Mel-only approach.

From a learning-dynamics perspective, MCQT achieved 30% lower final loss with stable convergence across folds, particularly on harmonically complex datasets. However, CQT's focused spectral representation occasionally outperforms MCQT in CMLt metrics (Table 3), suggesting that dual-input fusion can introduce representational noise that dilutes CQT's inherent spectral clarity when musical content aligns well with its harmonic-centric design. This trade-off indicates that fusion complexity may not always benefit precision-critical contexts.

Table 4: Academic Evaluation Scope (Phase Two) – Annotation Coverage Ratio (ACR)

Model	Annotation Coverage Ratio (ACR)								
	Any	Onbeat	Offbeat	Subharmonic Error			Harmonic Error		
				Half	Third	Quarter	Double	Triple	Quadruple
MEL	0.910	0.769	0.018	0.066	<0.001	<0.001	0.052	**0.007**	<0.001
CQT	0.905	0.779	0.019	0.059	**0.000**	0.000	**0.041**	0.010	<0.001
MCQT	**0.912**	**0.783**	0.019	**0.054**	0.001	<0.001	0.046	0.010	**0.000**

All values are reported as means across 8-fold cross-validation.
Values < 0.001 are shown as "<0.001" for readability.

Table 5: Production Evaluation Scope (Phase Two) – Ensemble-Based Analyses of Genre Classification Performance.

Genre	Mel Ensemble			CQT Ensemble			MCQT Ensemble		
	F-measure	CMLt	AMLt	F-measure	CMLt	AMLt	F-measure	CMLt	AMLt
Blues	*0.846*	*0.747*	*0.849*	**0.854**	**0.753**	**0.856**	0.842	0.720	0.846
Classical	*0.635*	**0.481**	0.651	*0.635*	**0.481**	*0.660*	**0.644**	**0.481**	**0.674**
Country	*0.926*	0.811	*0.960*	*0.926*	*0.818*	0.939	**0.934**	**0.835**	**0.961**
Disco	0.964	*0.943*	*0.967*	*0.965*	0.946	0.970	**0.968**	**0.948**	**0.972**
Hip-hop	*0.978*	*0.975*	*0.982*	**0.979**	**0.978**	**0.988**	0.968	0.953	0.970
Jazz	*0.872*	0.766	*0.880*	**0.874**	*0.769*	0.866	*0.872*	**0.775**	**0.881**
Metal	**0.882**	*0.764*	**0.884**	0.876	**0.792**	*0.863*	*0.879*	0.760	0.854
Pop	*0.952*	*0.908*	*0.960*	0.951	*0.908*	0.954	**0.961**	**0.930**	**0.970**
Reggae	0.876	0.649	0.945	*0.888*	*0.680*	*0.953*	**0.895**	**0.700**	**0.961**
Rock	**0.931**	**0.870**	0.921	0.899	0.839	*0.924*	*0.925*	**0.870**	**0.933**

Finally, the cross-validation ACR and ensemble genre-wise analyses on GTZAN reveal that MCQT's dual-branch pipeline adapts flexibly to both percussive and highly harmonic styles, achieving lower subharmonic error rates than single-representation baselines and superior performance across most of genres. This resilience indicates that MCQT can serve as a universal front-end for sequence-to-sequence beat trackers deployed in diverse musical environments, though practitioners should consider the trade-off between architectural complexity and the focused precision that single-representation models like CQT can provide in specialized applications. Music Information Retrieval ()

6. Conclusions

We have shown that integrating spectrogram representations with complementary time-frequency characteristics—namely, the fixed-window, perceptually motivated Mel scale and the logarithmic, pitch-aligned Constant-Q Transform—yields substantial gains in beat and downbeat tracking accuracy. The proposed MCQT architecture outperforms Mel-only and CQT-only models, accelerates learning convergence, and generalizes more robustly to unseen genres, including complex jazz and percussion-heavy corpora.

These findings underscore the often-overlooked role of input representation in MIR pipelines: rather than solely focusing on deeper or wider network architectures, researchers and practitioners should consider hybrid spectral embeddings as a powerful lever for performance improvement. Future work can explore to extend our framework to other sequence modeling tasks such as chord recognition or structural segmentation.

References

[1] Anastasia Natsiou and Sean O'Leary. Audio representations for deep learning in sound synthesis: a review. *arXiv preprint arXiv:2201.02490*, 2022.

[2] Friedrich Wolf-Monheim. Spectral and rhythm features for audio classification with deep convolutional neural networks, 2024.

[3] Judith C. Brown. Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

[4] Christian Schörkhuber and Anssi Klapuri. Constant-q transform toolbox for music processing. In *7th Sound and Music Computing Conference, Barcelona, Spain*, page 3, 2010.

[5] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[6] Sebastian Böck, Florian Krebs, and Gerhard Widmer. Joint beat and downbeat tracking with recurrent neural networks. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, pages 255–261. New York, NY, USA, 2016.

[7] Rainer Kelz, Sebastian Böck, and Gerhard Widmer. An end-to-end neural network for polyphonic piano music transcription. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, pages 676–682. New York, NY, USA, 2016.

[8] Yicheng Gu, Xueyao Zhang, Liumeng Xue, and Zhizheng Wu. Multi-scale sub-band constant-q transform discriminator for high-fidelity vocoder, 2023.

[9] Hongming Guo, Ruibo Fu, Yizhong Geng, Shuai Liu, Shuchen Shi, Tao Wang, Chunyu Qiang, Chenxing Li, Ya Li, Zhengqi Wen, Yukun Liu, and Xuefei Liu. Mel-refine: A plug-and-play approach to refine mel-spectrogram in audio generation, 2024.

[10] Micael Antunes da Silva. *Explorando Texturas Sonoras: Análise e Composição com Suporte de Descritores de Áudio e Modelos Computacionais*. Tese (doutorado), Universidade Estadual de Campinas (UNICAMP), Instituto de Artes, Campinas, SP, 2024. Disponível em: 20.500.12733/20336. Acesso em: 14 jun. 2025.

[11] Francesco Foscarin, Jan Schlüter, and Gerhard Widmer. Beat this! accurate beat tracking without dbn postprocessing. `https://arxiv.org/abs/2407.21658`, 2024.

[12] Haytham Fayek. Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients (mfccs) and what's in-between. `https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html`, April 2016.

[13] Simon Dixon. Automatic music transcription. In *Music Data Mining*, pages 267–294. CRC Press, 2007.

[14] Sebastian Böck and Markus Schedl. Enhanced beat tracking with context-aware neural networks. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 657–660. ACM, 2012.

[15] Joaquim B. Cavalcante and Julio Hsu. Beat this: Multi-representation beat tracking implementation. `https://github.com/JoaquimBreno/rep_matters_beat`, 2025.

[16] Jon Gillick, Adam Roberts, Jesse Engel, Douglas Eck, and David Bamman. Learning to groove with inverse sequence transformations. In *International Conference on Machine Learning (ICML)*, volume 97, pages 2269–2279, 2019.

[17] Fabien Gouyon. *A computational approach to rhythm description—Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing*. PhD thesis, Universitat Pompeu Fabra, 2005.

[18] Jason A Hockman, Matthew EP Davies, and Ichiro Fujinaga. One in the jungle: Downbeat detection in hardcore, jungle, and drum and bass. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2012.

[19] Stephen W Hainsworth and Malcolm D Macleod. Particle filtering applied to musical tempo tracking. In *EURASIP Journal on Advances in Signal Processing*, volume 2004, 2004.

[20] Andre Holzapfel, Matthew EP Davies, José R Zapata, João Lobato Oliveira, and Fabien Gouyon. Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9), 2012.

[21] Vasiliy Eremenko, Eren Demirel, Barış Bozkurt, and Xavier Serra. Audio-aligned jazz harmony dataset for automatic chord transcription and corpus-based research. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2018.

[22] Jonathan Driedger, Hendrik Schreiber, W Bas De Haas, and Meinard Müller. Towards automatically correcting tapped beat annotations for music recordings. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[23] Dave Foster and Simon Dixon. Filosax: A dataset of annotated jazz saxophone recordings. In *International Society for Music Information Retrieval (ISMIR) Conference*, 2021.

[24] Qingyang Xi, Rachel M Bittner, Johan Pauwels, Xuzhou Ye, and Juan Pablo Bello. *GuitarSet: A dataset for guitar transcription*. PhD thesis, New York University, 2018.

[25] Matthew EP Davies, Norberto Degara, and Mark D Plumbley. Evaluation methods for musical audio beat tracking algorithms. Technical report, Queen Mary University of London, 2009.

[26] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database: Popular, classical and jazz music databases. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2002.

[27] Florian Krebs, Sebastian Böck, and Gerhard Widmer. Rhythmic pattern modeling for beat and downbeat tracking in musical audio. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2013.

[28] Francesco Foscarin, Andrew Mcleod, Philippe Rigaux, Florent Jacquemard, and Masahiko Sakai. Asap: a dataset of aligned scores and performances for piano transcription. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[29] Oriol Nieto, Matthew McCallum, Matthew Davies, Andrew Robertson, Adam Stark, and Eran Egozy. The harmonix set: Beats, downbeats, and functional segment annotations of western popular music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[30] Martín Rocamora, Luis Jure, et al. Beat and downbeat tracking based on rhythmic patterns applied to the uruguayan candombe drumming. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[31] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2019.

[32] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[33] George Tzanetakis, Georg Essl, and Perry Cook. Musical genre classification of audio signals, 2001.

[34] Matthew EP Davies and Mark D Plumbley. Evaluation of the audio beat tracking system beatroot. In *Journal of New Music Research*, volume 38, pages 117–129, 2009.

[35] Ching-Yu Chiu, Meinard Müller, Matthew E. P. Davies, Alvin Wen-Yu Su, and Yi-Hsuan Yang. An analysis method for metric-level switching in beat tracking. *IEEE Signal Processing Letters*, 29:2153–2157, 2022.

[36] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel PW Ellis. mir_eval: A transparent implementation of common mir metrics. *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pages 367–372, 2014.