Advancing Multi-Instrument Music Transcription: Results from the 2025 AMT Challenge

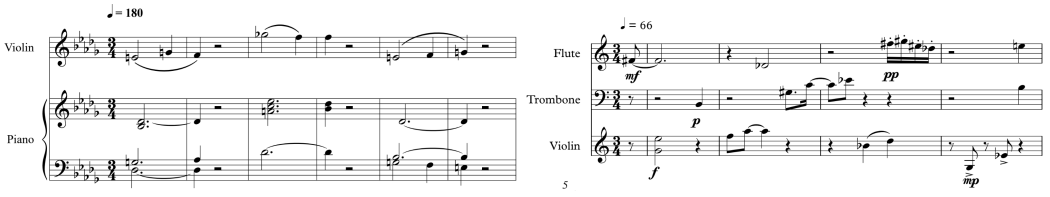
Ojas Chaturvedi, Kayshav Bhardwaj, Tanay Gondil, Benjamin Shiue-Hal Chou,
{ochaturv, bhardw43, tgondil, chou150}@purdue.edu
Kristen Yeon-Ji Yun, Yung-Hsiang Lu, Yujia Yan, Sungkyun Chang
{yun98, yunglu}@purdue.edu, yujia.yan@rochester.edu, sungkyun.chang@qmul.ac.uk

Abstract

This paper presents the results of the 2025 Automatic Music Transcription (AMT) Challenge, an online competition to benchmark progress in multi-instrument transcription. Eight teams submitted valid solutions; two outperformed the baseline MT3 model. The results highlight both advances in transcription accuracy and the remaining difficulties in handling polyphony and timbre variation. We conclude with directions for future challenges: broader genre coverage and stronger emphasis on instrument detection.

1 Introduction

Automatic Music Transcription (AMT) converts audio signals into symbolic representations of music, such as sheet music or MIDI (Musical Instrument Digital Interface) format. Compared to other fields of artificial intelligence (such as natural language processing and computer vision), the progress of AMT has been slower [4]. Several factors contribute to this gap: (1) lack of large openly available datasets for training and evaluation, (2) absence of commonly adopted benchmarks for comparing different methods, (3) limited incentives to attract researchers compared to other fields in AI, (4) lack of a common platform where multiple AMT solutions can be evaluated., and (5) the inherent complexity of music signals. A single instrument can produce wide variations in pitch, articulation, dynamics, and playing techniques. Note onsets/offsets themselves can be ambiguous; e.g., piano notes may be defined acoustically (sounding strings) or by key press and release with or without pedals. In this work, we adopt the `mir_eval` convention [16], where onsets are defined by the reference MIDI start time and offsets by the later of 50 ms or 20% of note duration. This paper presents the results of the 2025 Automatic Music Transcription Challenge, designed to advance AMT technology by fostering benchmarking, comparison, and community engagement.



(a) (b)

Figure 1: Excerpts of the sample music available to the participants.

This challenge is different from previous competitions in several ways: (1) This challenge offers a cloud-based grading system that updates the leaderboard on a daily basis. Such timely information encourages participants to improve their methods progressively. (2) The evaluation metrics include

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

testing for multiple factors, as well as taking into account tolerances in pitch. (3) To avoid overfitting to existing datasets, the challenge introduced a newly composed test set comprising 76 short pieces (≈20 seconds each) across eight instruments; up to three instruments may be present in each piece. Figure 1 shows two excerpts of the released music. Each piece includes three files: (1) a PDF file of the sheet music, (2) an MP3 of synthesized audio, and (3) a MIDI file.

2 Previous Transcription Competitions

Music transcription has been evaluated through a variety of competitions, each designed to address specific aspects of the task and to drive progress. Since its inception in 2005, the Music Information Retrieval Evaluation eXchange (MIREX) [3] has served as a pivotal benchmark, initially focusing on foundational tasks such as onset detection, tempo and beat tracking, melody extraction, key detection, and chord estimation, with submissions typically targeting one component of the broader transcription pipeline. Over time, challenges evolved toward instrument-specific, polyphonic settings, notably the Drum Transcription challenge [1] and the Polyphonic Piano Transcription challenge [2]. The former emphasized detecting bass drum, snare drum, and hi-hat in polyphonic mixes; the latter (introduced in 2024) required systems to convert solo piano audio into MIDI by capturing onsets, offsets, pitch, and velocity, leveraging the MAESTRO v3.0.0 dataset for training and evaluation.

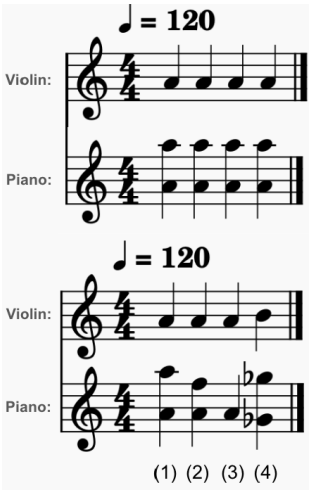


Figure 2: Scoring Method. Top: Reference. Bottom: Estimated transcription.

3 Evaluation Methodology

3.1 Scoring System

Two MIDI files are compared to calculate the transcription program's score. The evaluation metrics include precision, recall, F1 score, and overlap ratio. The precision and recall of the reference and estimated MIDI are computed using the mir_eval library [16]. Overlap is calculated through onset and offset, i.e., the timing of a music note's beginning and ending, by computing the intersection over union between transcription and ground truth.

Consider Figure 2. The top is the reference (correct). The bottom is the estimated (transcription). The excerpts are divided into four groups. Each group is one quarter note long, and each group has a slight variation from the reference. Group (1) has the reference and estimated tracks as identical and receives an F1 score of 1.0. Group (2) has the top note in the piano part as incorrect. It correctly finds two notes, but misses a correct one, which is counted as a false negative. The wrong note is counted as a false positive. Using the formulas, precision and recall are both 0.667, leading to group (2) and receiving an F1 score of 0.667. Group (3) has the top note of the piano part as completely missing. Therefore, it has two true positives for the correct note, one false negative for the missing note, and no false positives. Precision is 1.00, and recall is 0.667. Group (3) receives an F1 score of 0.800. Group (4) has all three notes from both instruments incorrect. It has three false negatives for the missing true positive notes and three false positives for the wrong notes, with no true positives, causing precision and recall to be 0. Group (4) receives an F1 score of 0.000. The overall F1 score of the entire excerpt is 0.609.

3.2 Test Data

To fairly benchmark transcription models, the challenge used a new evaluation set of 76 pieces written by five professional composers, including for the first time newly composed modern atonal works and rare coverage of instruments such as bassoon and viola. These instruments sparsely in existing datasets like MusicNet and URMP. The audio was rendered directly from the MIDI scores using FluidSynth with the FluidR3 GM soundfont. Each composition followed a set of predefined rules announced to all participants: (1) Tempi restricted to 60–90 BPM; (2) Meters limited to 3/4, 4/4, or 6/8; (3) Maximum rhythmic subdivision of sixteenth notes; (4) Richer notational elements such as swing, double-dotted

notes, grace notes, and trills were excluded to reduce ambiguity in alignment; (5) Pitch range limited to C2–C7 (five octaves); (6) Dynamic markings restricted between pianissimo and fortissimo; (7) Eight instruments were allowed (Piano, Violin, Viola, Cello, Flute, Bassoon, Trombone, Oboe), with at most three instruments per piece; (8) At most one string instrument could appear in a piece. These constraints were chosen to balance *musical realism* (e.g., polyphony, dynamics, instrument variety) with *evaluation clarity*, avoiding edge cases where subjective interpretation or notational ambiguity might dominate scoring. The constraints were also applied to simplify the task for the competitors, and to ensure the model's scoring remained sufficiently high.

4 Competition Results

A total of 21 teams registered for the 2025 Automatic Music Transcription Challenge, of which 14 teams submitted at least one solution. Eight teams submitted valid solutions whose MIDI outputs could be successfully graded without errors. Table 1 reports the results across all evaluated models, including open-source baseline (MT3). To protect privacy, teams can choose their own names. In the discussion, we focus on the top three systems, as lower-ranked submissions showed limited improvements over the baseline.

Table 1: Results from all evaluated systems. Multiple variants and public baselines are included; MT3 serves as the reference baseline. Runtime is measured by ms.

Rank	Model Name	F1 Score	Precision	Recall	Overlap	Runtime
1	MIROS	**0.5998**	**0.6558**	0.5724	0.7391	22.05
2	YourMT3-YPTF-MoE-M	0.5938	0.6010	**0.5888**	0.7305	12.60
3	YourMT3-YPTF-S	0.5581	0.5565	0.5615	0.7326	15.40
4	YourMT3-P	0.3947	0.3966	0.3985	0.7263	14.99
5	MT3 [11] (baseline)	0.3932	0.3811	0.4115	0.7180	20.19
6	YourMT3-YPTF-SP-V	0.3305	0.3280	0.3358	0.7147	14.50
7	press_to_win 1	0.3199	0.3105	0.3346	0.7331	19.30
8	press_to_win 2	0.3190	0.3094	0.3331	0.7310	18.08
9	YourMT3-YPTF-MoE-MP	0.2173	0.2150	0.2206	0.6116	16.03
10	press_to_win 3	0.2168	0.2144	0.2203	0.6159	16.15
11	Bytedance Piano [13]	0.1721	0.2041	0.1689	0.5423	9.67
12	press_to_win 4	0.1470	0.1305	0.1799	0.6998	21.74
13	ReconVAT [8]	0.1415	0.1215	0.1803	**0.7898**	5.45
14	Basic Pitch [5]	0.0634	0.0550	0.0782	0.5977	**3.91**

Music Information Retrieval Osnabrück (Winning System). Music Information Retrieval Osnabrück (MIROS) extends the YourMT3+ encoder–decoder framework for automatic music transcription by adopting *MusicFM* as the encoder backbone and pairing it with modernized decoders. MusicFM is a conformer-based, self-supervised foundation model for music pretrained via BEST-RQ masked token modeling [17]. An advantage of self-supervised foundation models is that they can exploit abundant unlabeled audio, compared with scarce labeled AMT data. To integrate MusicFM into the YourMT3+ multi-decoder formulation, MIROS introduced a recurrent adapter that conditions the temporally downsampled encoder outputs on learned instrument group embeddings. Each instrument group is then decoded in parallel using T5-style decoders with cross attention, updated with rotary position embeddings and hardware-optimized attention (FlashAttention) [10]. Unlike prior YourMT3+ state-of-the-art systems, MIROS did not employ cross-stem augmentation, though they did train on all available MIDI datasets [6]. This isolates the contribution of domain-pretrained audio representations within a comparable seq2seq framework.

Outside the competition, the MusicFM + multi-decoder system ($\approx$370M parameters) achieved a Slakh2100 [15] multi-instrument F-measure of 0.83, with efficient long-context decoding (5-second windows, up to 1024 tokens per group) and high training throughput ($\approx$2.4 iterations/s). Although it underperformed YourMT3+ on Slakh2100, it attained slightly better accuracy on the competition data, suggesting possible Slakh overfitting by YourMT3+. Domain-pretrained encoders like MusicFM thus extends YourMT3+ to longer, richer contexts while maintaining competitive accuracy, while optimized attention backends improve the practicality of multi-instrument AMT at scale.

General Trends Across Models. Several themes emerged from the technical directions of top models. Most submissions adopted a sequence-to-sequence (Seq2Seq) paradigm, extending the MT3 [11] architecture with enhancements such as self-supervised random projection quantizer [17, 9], Mixture-of-Experts [7, 12] routing, hierarchical time–frequency attention [14], cross-dataset augmentation [7], and auxiliary onset/offset losses. A second theme was efficiency: some teams targeted real-time transcription by pruning parameters, quantizing model weights, and lowering spectrogram resolution to reduce GPU memory and inference latency. Despite these advances, common failure cases were observed. Models often produced *instrument leakage*, hallucinating nonexistent instruments, or struggled to disentangle salient melodies from dense polyphonic textures, especially when multiple instruments shared similar pitch ranges or timbres.

Since instrument leakage and polyphonic confusion emerged as common failure modes, we conducted a focused statistical analysis on transcription accuracy in single-instrument versus multi-instrument settings. A two-way ANOVA on `f_measure` with factors *model* and *instrument category* confirmed that the number of input instruments had a strong main effect ($F(1, 148) = 27.5$, $p < 0.001$), while the model main effect ($p = 0.15$) and the interaction ($p = 0.43$) were not significant. This indicates that all systems consistently perform worse on multi-instrument tracks, regardless of architecture.

To expand this further, we conducted pairwise Welch's two-sample t-tests comparing single- versus multi-instrument excerpts within each model. For the top system (A), the difference was large in magnitude ($t = 2.84$, $p = 0.0057$, Cohen's $d = 1.21$). For YourMT3-YPTF-MoE-M, the difference was even more pronounced ($t = 4.84$, $p < 0.00001$, Cohen's $d = 2.06$). Because multiple comparisons were performed across models, we also applied a Bonferroni correction. After correction, the difference for MIROS became borderline significant ($p \approx 0.055$), while the YourMT3-YPTF-MoE-M result remained robustly significant ($p = 0.024$). We report both raw and corrected p-values to distinguish statistical from practical significance. In both cases, the effect sizes exceeded $d = 1.2$, which corresponds to very large practical differences.

Table 2 provides descriptive statistics for the top two systems, broken down by single- and multi-instrument excerpts. Both systems show a consistent drop of more than 0.25 F-measure points on multi-instrument settings, with MIROS exhibiting especially sharp precision loss.

Table 2: Performance comparison between single- and multi-instrument for the top two teams. Values are reported as mean ± standard deviation. Welch's t-test results are shown above in the text.

Model	Instrument Type	F-measure	Precision	Recall
MIROS	Multiple (n=70)	0.4585 ± 0.2158	0.4923 ± 0.2153	0.4424 ± 0.2251
MIROS	Single (n=6)	$\mathbf{0.7193 \pm 0.2103}$	$\mathbf{0.9067 \pm 0.2286}$	$\mathbf{0.6233 \pm 0.2295}$
YourMT3-YPTF-MoE-M	Multiple (n=70)	0.4055 ± 0.1668	0.4128 ± 0.1696	0.3998 ± 0.1653
YourMT3-YPTF-MoE-M	Single (n=6)	$\mathbf{0.7594 \pm 0.2304}$	$\mathbf{0.7858 \pm 0.2411}$	$\mathbf{0.7367 \pm 0.2254}$

Overall, nearly all top-performing models and the MT3 baseline relied on an almost identical set of ten datasets [11, 7]. Notably, both YourMT3-P and MT3 applied cross-dataset augmentation on the same architecture and training corpus, yet achieved only marginal improvements. This suggests that performance is fundamentally constrained by data scarcity. In particular, currently available public datasets for AMT insufficiently cover less common instruments such as viola, bassoon, and trombone, and are largely limited to a small number of ensemble pieces, thereby restricting generalization.

5 Conclusion and Future Directions

The 2025 Automatic Music Transcription Challenge advanced transcription of multi-instrument music by introducing a new test set and a cloud-based evaluation system. Several submissions surpassed the MT3 baseline, showing tangible gains but also exposing persistent weaknesses: dense polyphony, timbrally similar instruments, and limited data diversity. These limitations suggest that progress in AMT will depend as much on richer training resources as on architectural innovation. Future iterations will expand to genres such as jazz and popular music and emphasize robust instrument detection. Additional improvements will target evaluation protocols that better capture pitch, timing, and timbre quality. By broadening coverage and refining benchmarks, future challenges aim to support the development of systems that generalize more reliably to diverse music.

References

[1] Drum Transcription, Sept. 2021. URL https://www.music-ir.org/mirex/wiki/2021:Drum_Transcription.

[2] Polyphonic Transcription, Oct. 2024. URL https://www.music-ir.org/mirex/wiki/2024:Polyphonic_Transcription.

[3] MIREX Home, 2025. URL https://www.music-ir.org/mirex/wiki/MIREX_HOME.

[4] E. Benetos, S. Dixon, Z. Duan, and S. Ewert. Automatic Music Transcription: An Overview. *IEEE Signal Processing Magazine*, 36(1):20–30, Jan. 2019. ISSN 1558-0792. doi: 10.1109/MSP.2018.2869928. URL https://ieeexplore.ieee.org/document/8588423. Conference Name: IEEE Signal Processing Magazine.

[5] R. M. Bittner, J. J. Bosch, D. Rubinstein, G. Meseguer-Brocal, and S. Ewert. A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Singapore, 2022.

[6] S. Chang, S. Dixon, and E. Benetos. YourMT3: A toolkit for training multi-task and multi-track music transcription model for everyone, Dec. 2022.

[7] S. Chang, E. Benetos, H. Kirchhoff, and S. Dixon. Yourmt3+: Multi-instrument music transcription with enhanced transformer architectures and cross-dataset stem augmentation. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2024.

[8] K. W. Cheuk, D. Herremans, and L. Su. ReconVAT: A Semi-Supervised Automatic Music Transcription Framework for Low-Resource Real-World Data. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pages 3918–3926, New York, NY, USA, Oct. 2021. Association for Computing Machinery. ISBN 978-1-4503-8651-7. doi: 10.1145/3474085.3475405. URL https://dl.acm.org/doi/10.1145/3474085.3475405.

[9] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu. Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning*, pages 3915–3924. PMLR, 2022.

[10] T. Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023.

[11] J. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. Engel. MT3: Multi-Task Multitrack Music Transcription, Mar. 2022. URL http://arxiv.org/abs/2111.03017. arXiv:2111.03017 [cs].

[12] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[13] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang. High-Resolution Piano Transcription With Pedals by Regressing Onset and Offset Times. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3707–3717, 2021. ISSN 2329-9304. doi: 10.1109/TASLP.2021.3121991. URL https://ieeexplore.ieee.org/abstract/document/9585550.

[14] W.-T. Lu, J.-C. Wang, and Y.-N. Hung. Multitrack music transcription with a time-frequency perceiver. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[15] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019.

[16] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis. mir_eval: A Transparent Implementation of Common MIR Metrics. In *Proceedings of the 15th International Conference on Music Information Retrieval (ISMIR 2014)*, pages 367–372, 2014. URL https://archives.ismir.net/ismir2014/paper/000320.pdf.

[17] M. Won, Y.-N. Hung, and D. Le. A foundation model for music informatics. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1226–1230. IEEE, 2024.