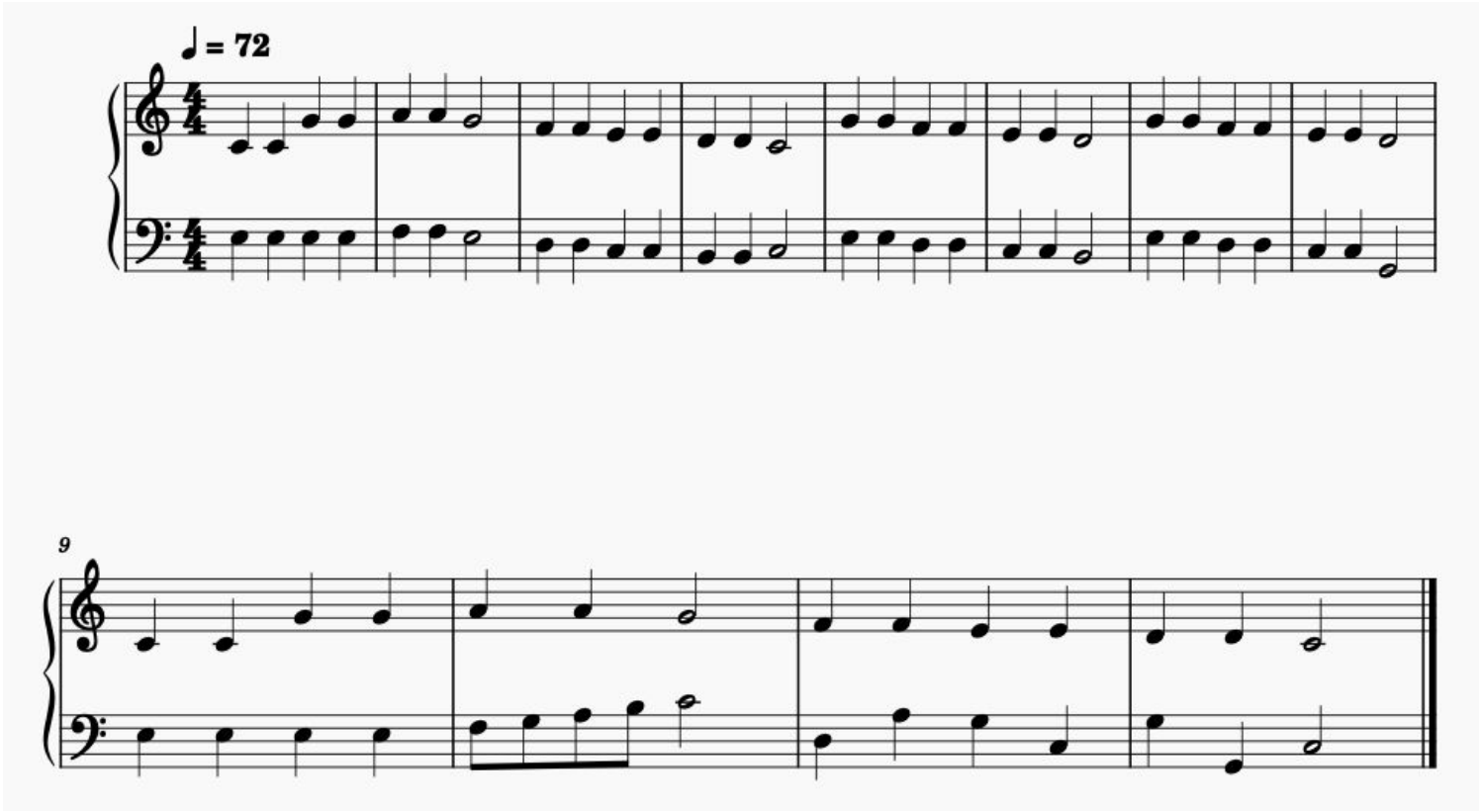
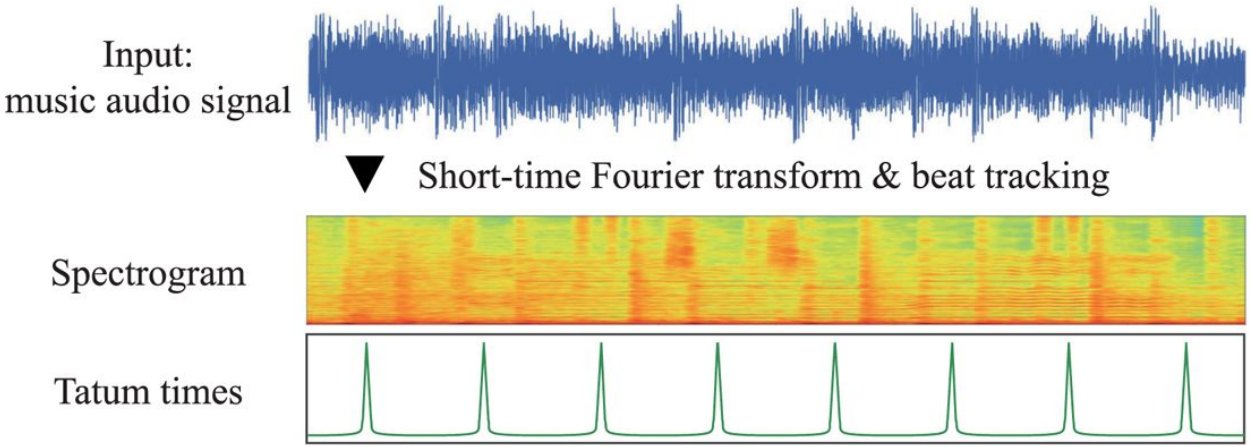
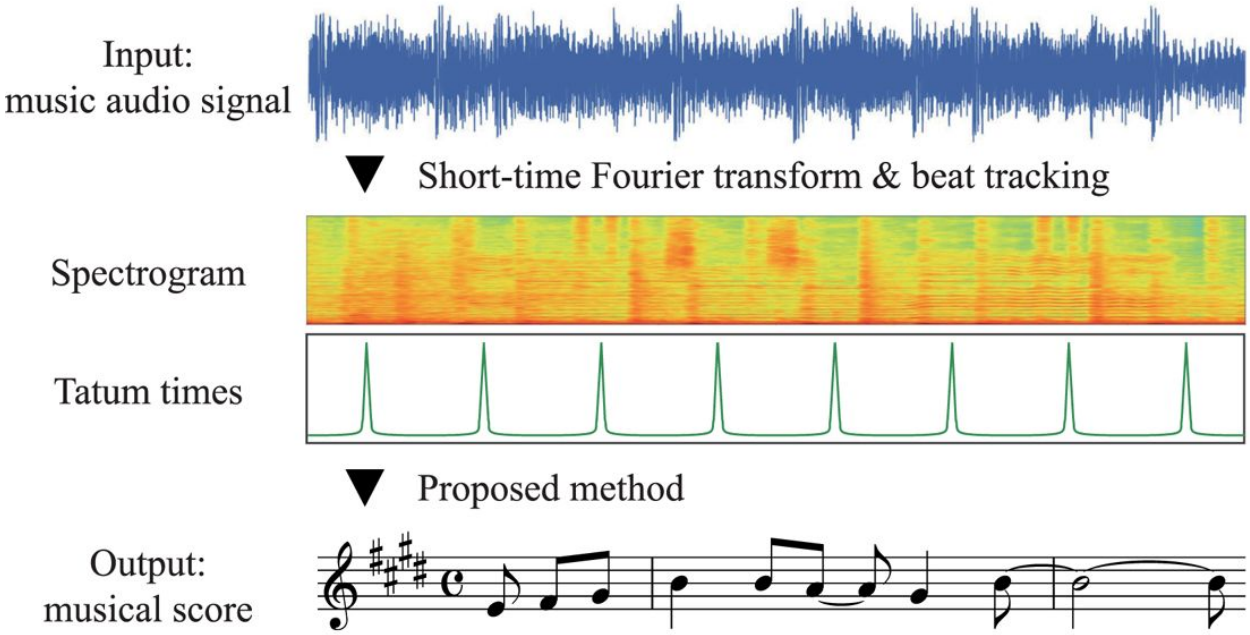
Automatic Music Transcription

Ojas Chaturvedi, Kayshav Bhardwaj, Tanay Gondil, Ritwik Jayaraman, Franklin Shang, Emily Li, Sean Su, Shreeya Sarurkar, Elliott Soderberg, Arnav Kalekar, Vishaal Iyer, Junyong Lee, Basil Khwaja
• AI4Musicians Lab


♩ = 72
9

Input:
music audio signal


Input:
music audio signal
Short-time Fourier transform & beat tracking
Spectrogram
Tatum times

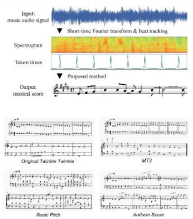

Input:
music audio signal
Short-time Fourier transform & beat tracking
Spectrogram
Tatum times
Proposed method
Output:
musical score

Past and Current Work

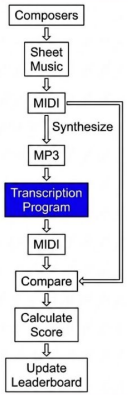
Advancing Multi-Instrument Music Transcription: Results from the 2025 AMT Challenge

Ojas Chaturvedi (ochaturv@purdue.edu), Kayshav Bhardwaj, Tanay Gondil, Benjamin Shiue-Hal Chou, Yujia Yan, Kristen Yeon-Ji Yun, Yung-Hsiang Lu, Sungkyun Chang

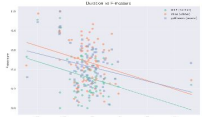
Current Field



Workflow



Performance Considerations

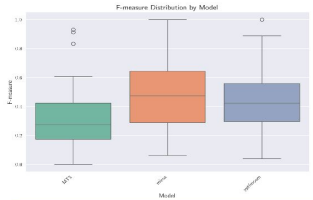


We observed two clear trends: model performance declined on longer pieces and on pieces with more overlapping instruments.

We launched the **2025 AMT Challenge** with a novel multi-instrument dataset, revealing a critical gap in polyphonic transcription.

Results

Rank	Model Name	F1 Score	Precision	Recall	Overlap	Runtime
1	MIROS	**0.5998**	**0.6588**	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 (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	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	0.1415	0.1215	0.1803	**0.7898**	5.45
14	Basic Pitch	0.0634	0.0550	0.0782	0.5977	**3.91**



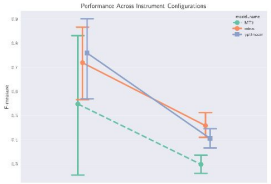
Architectures

1. MIROS

- Uses MusicFM, a conformer-based, self-supervised foundation model, which is pretrained via BEST-RQ masked token modeling to leverage abundant unlabeled audio data.
- Extends the YourMT3+ framework with parallel T5-style multi-decoders, utilizing rotary position embeddings and hardware-optimized FlashAttention.
- Showed better accuracy on the new competition data compared to systems potentially overfit to existing datasets.

2. YourMT3-YPTF-MoE-M

- Achieved competitive accuracy (F1 0.5938) with a significantly fast runtime (12.60 ms), demonstrating excellent practical efficiency.
- Achieved the highest Recall (0.5888) among all submissions, indicating superior success in detecting all true musical notes.
- Demonstrated robust foundational performance on single-instrument tracks (F1 0.7594).
- Part of a model family that incorporates enhancements like Mixture-of-Experts (MoE) routing within its sequence-to-sequence architecture.

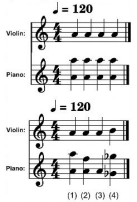


Dataset

The evaluation set consists of 76 newly composed pieces – including modern atonal works and underrepresented instruments such as bassoon and viola – rendered from MIDI using FluidSynth (FluidR3 GM). Each piece follows strict constraints: tempos between 60-90 BPM, meters limited to 3/4, 4/4, or 6/8, and rhythmic subdivisions no smaller than sixteenth notes. To avoid ambiguity, elements like swing, trills, double-dotted rhythms, and grace notes were excluded. Pieces span a pitch range of C2–C7, use dynamics from pp to ff, and include up to three instruments drawn from eight allowed choices, with at most one string instrument per piece.

Scoring

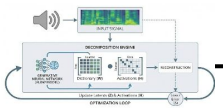
Two MIDI files are compared to calculate the transcription program's score. The evaluation metrics include precision, recall, F1 score, and overlap ratio. The precision and recall of the reference and estimated MIDI are computed using the mir_eval library. Overlap is calculated through onset and offset,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

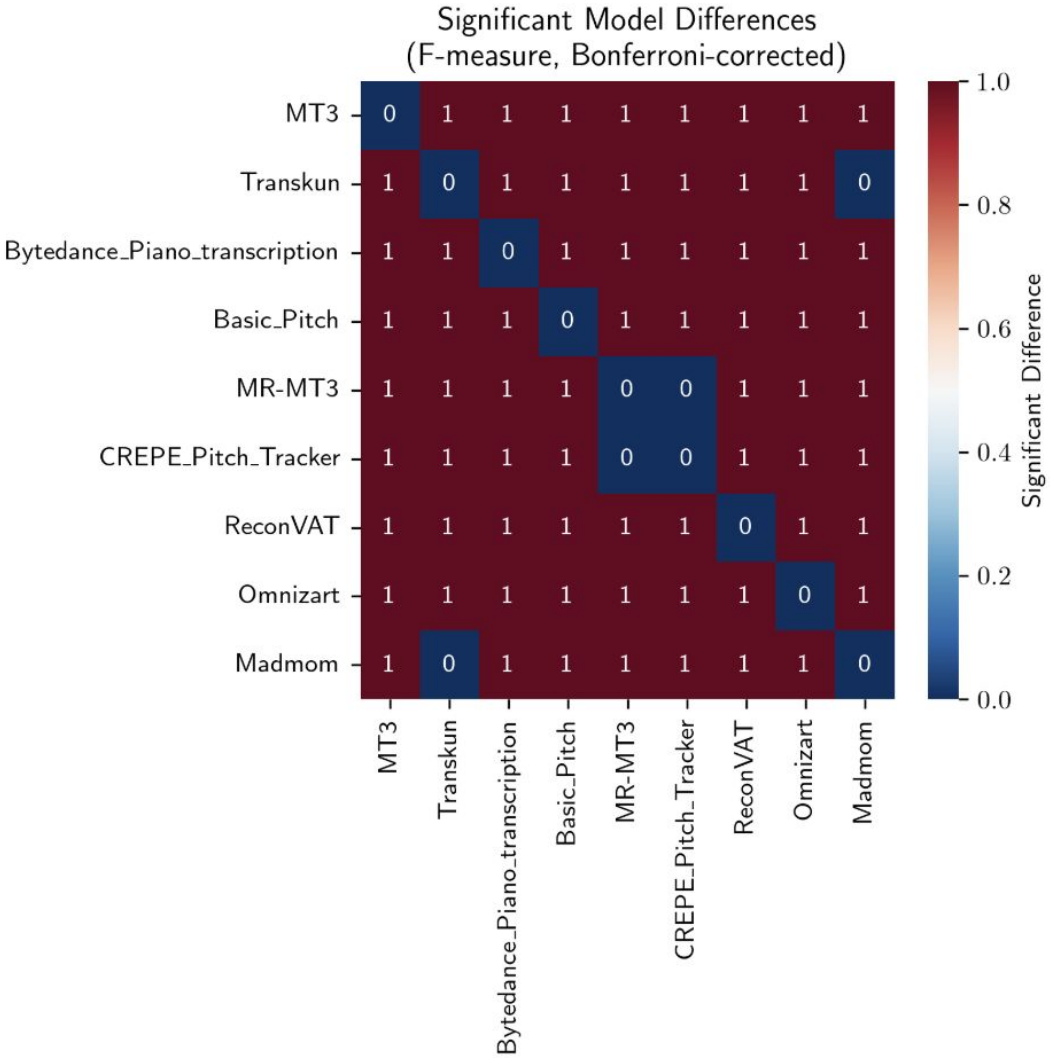


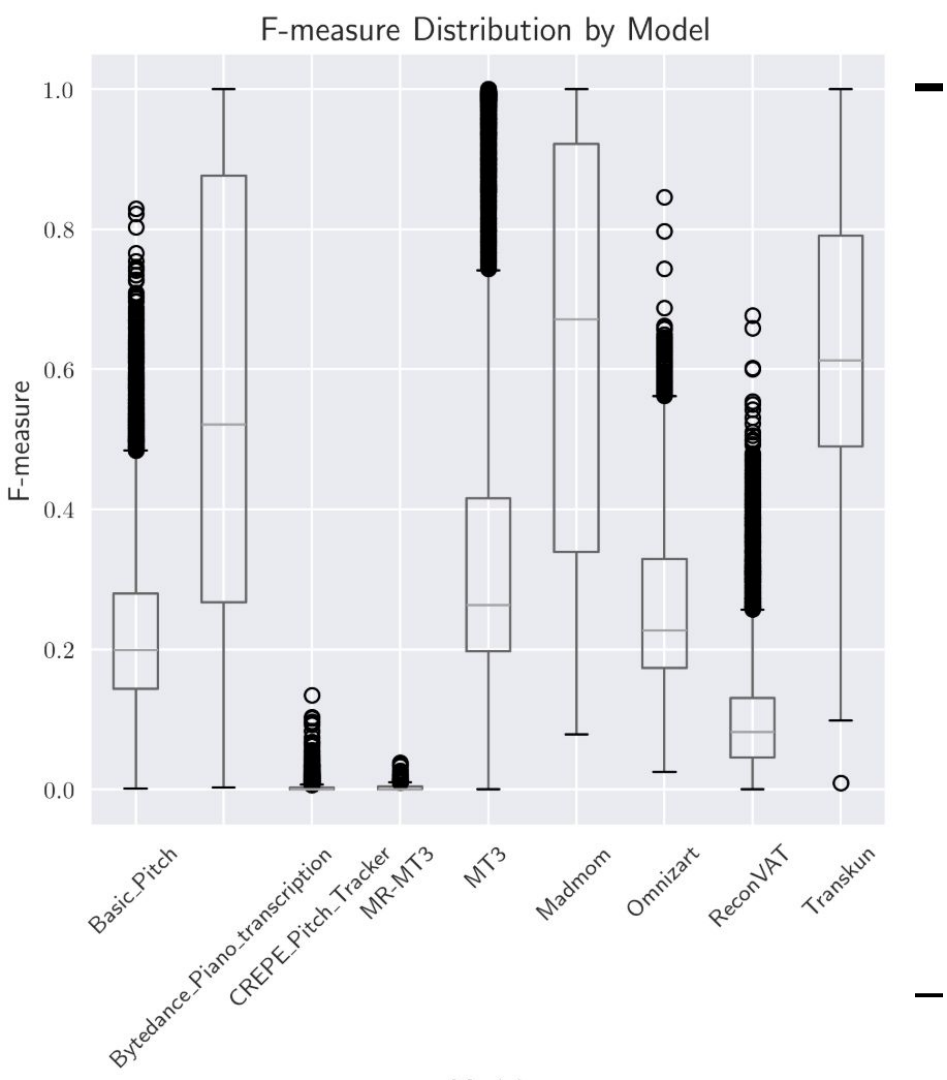
Future Work

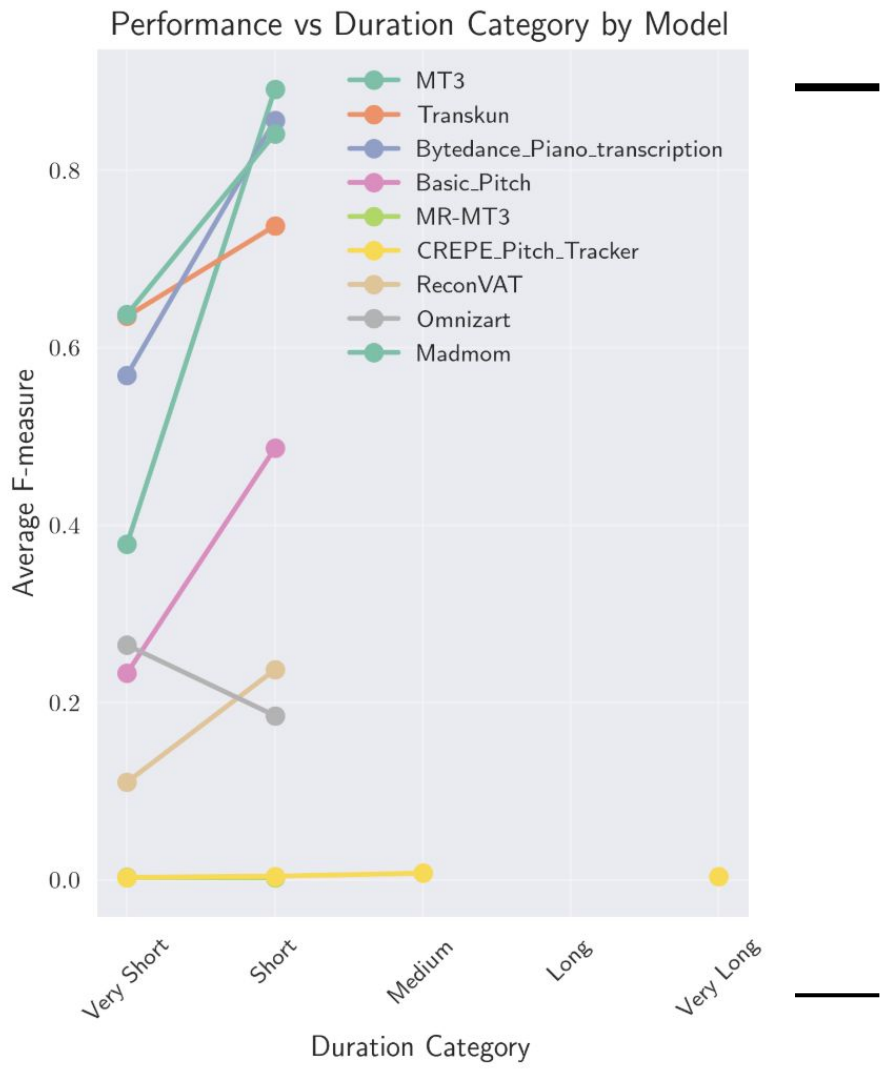
Our goal is to build an **interpretable** AMT model – one that approaches state-of-the-art accuracy while offering greater transparency. We aim to design a system whose internal decisions can be examined and understood, rather than treated as a black box.

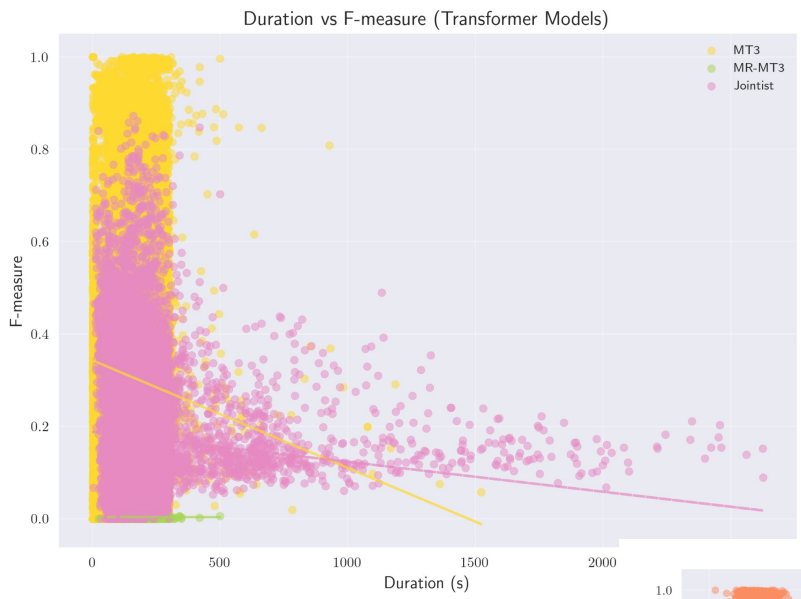
To achieve this, we plan to employ the DDS (Differentiable Dictionary Search) architecture. DDS uses a normalizing flow to learn a dictionary of possible "notes" and represents an audio signal as a linear combination of these learned bases. This allows the model not only to make accurate predictions but also to reveal which dictionary elements it relies on, providing a clearer window into how the transcription is produced.

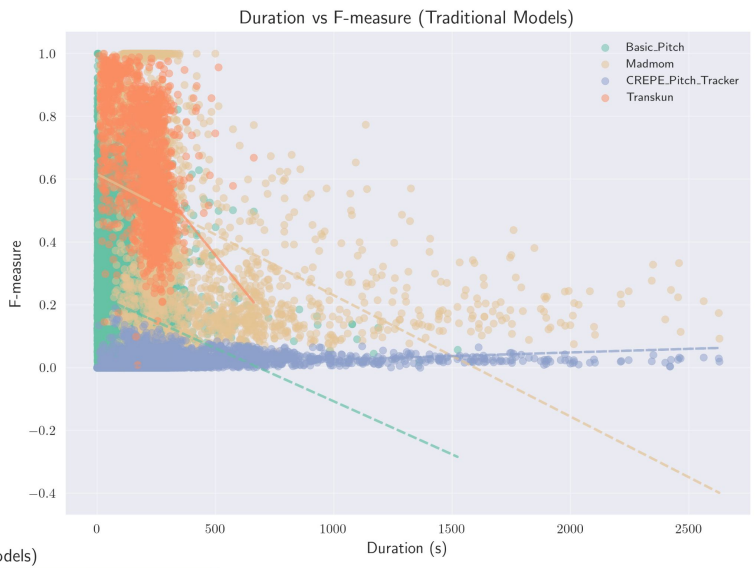


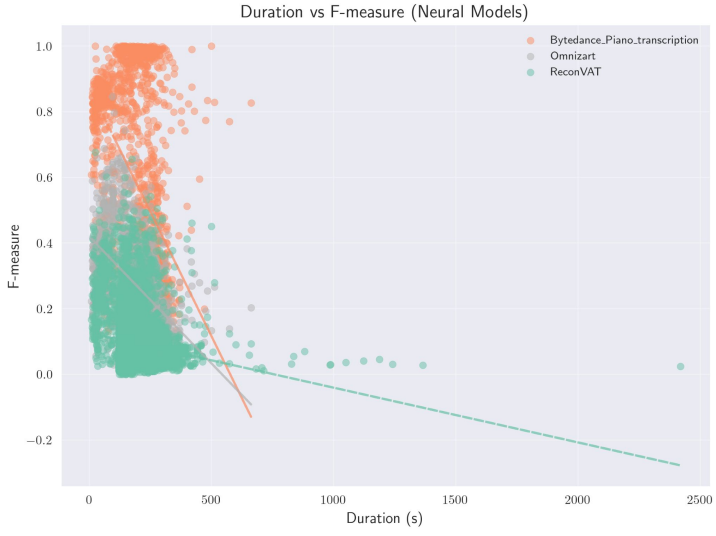

Significant Model Differences
(F-measure, Bonferroni-corrected)
MT3
Transkun
Bytedance_Piano_transcription
Basic_Pitch
MR-MT3
CREPE_Pitch_Tracker
ReconVAT
Omnizart
Madmom
Significant Difference
1.0
0.8
0.6
0.4
0.2
0.0

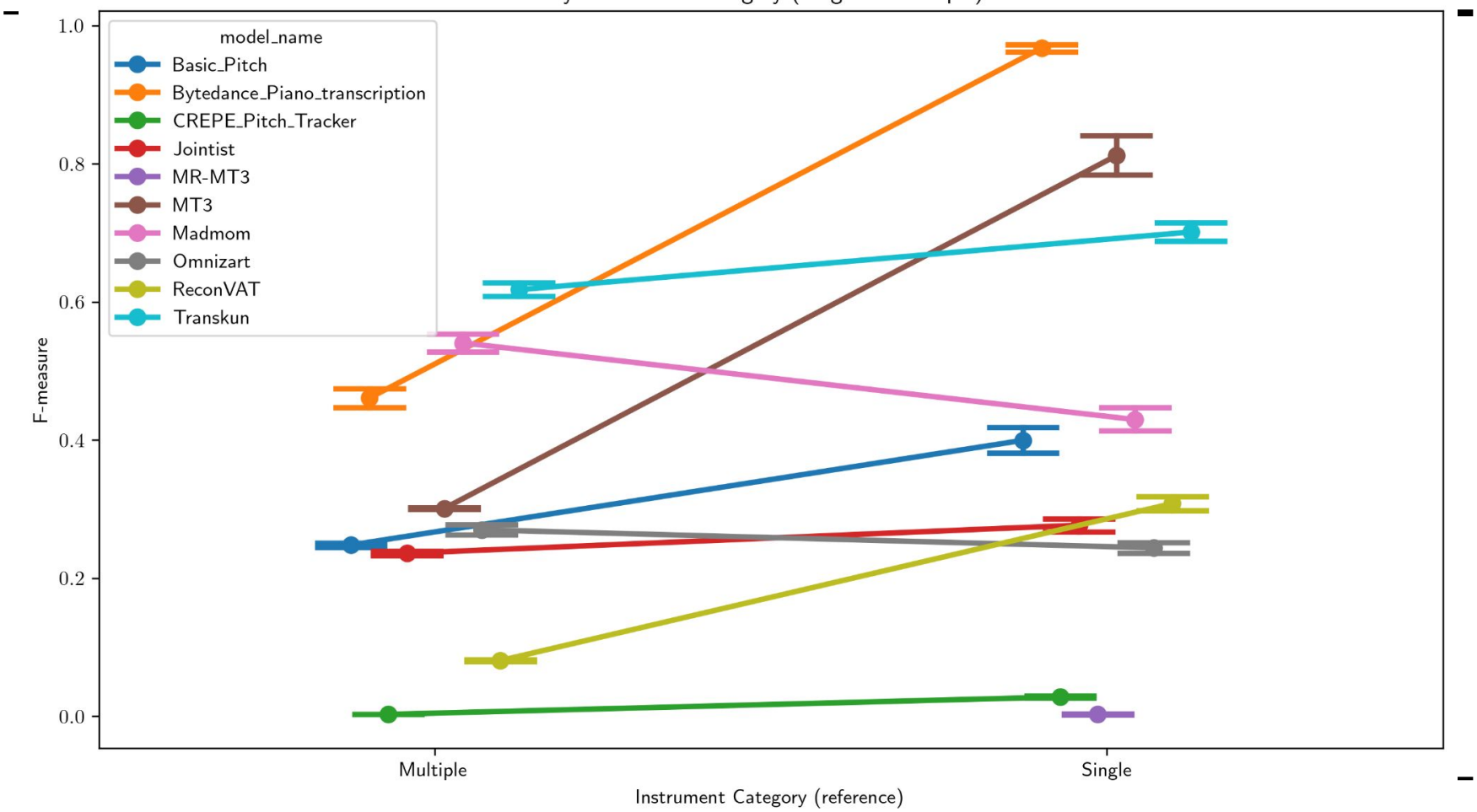

F-measure Distribution by Model
F-measure
1.0
0.8
0.6
0.4
0.2
0.0
Basic_Pitch
Bytedance_Piano_transcription
CREPE_Pitch_Tracker
MR-MT3
MT3
Madmom
Omnizart
ReconVAT
Transkun


Performance vs Duration Category by Model
Average F-measure
0.8
0.6
0.4
0.2
0.0
MT3
Transkun
Bytedance_Piano_transcription
Basic_Pitch
MR-MT3
CREPE_Pitch_Tracker
ReconVAT
Omnizart
Madmom
Very Short
Short
Medium
Long
Very Long
Duration Category

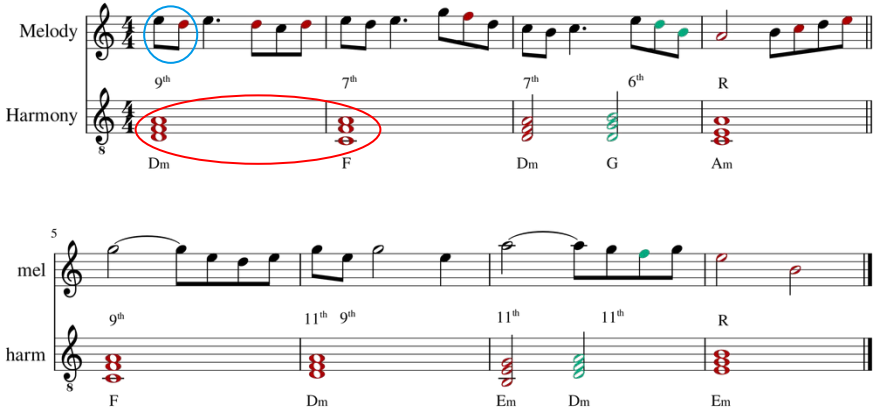

Duration vs F-measure (Transformer Models)
MT3
MR-MT3
Jointist
F-measure
Duration (s)


Duration vs F-measure (Traditional Models)
Basic_Pitch
Madmom
CREPE_Pitch_Tracker
Transkun
F-measure
Duration (s)


Duration vs F-measure (Neural Models)
Bytedance_Piano_transcription
Omnizart
ReconVAT
F-measure
Duration (s)

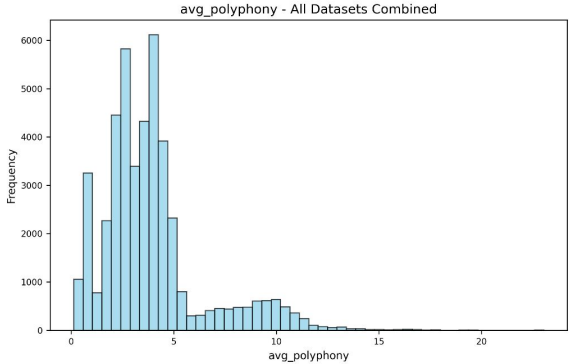

F-measure by Instrument Category (Single vs Multiple) and Model
model_name
Basic_Pitch
Bytedance_Piano_transcription
CREPE_Pitch_Tracker
Jointist
MR-MT3
MT3
Madmom
Omnizart
ReconVAT
Transkun
F-measure
1.0
0.8
0.6
0.4
0.2
0.0
Multiple
Single
Instrument Category (reference)

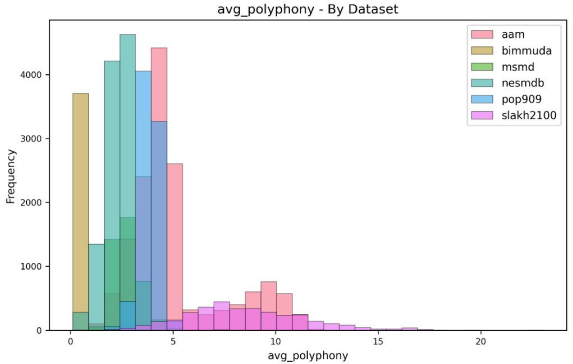
Music Complexity - Metrics

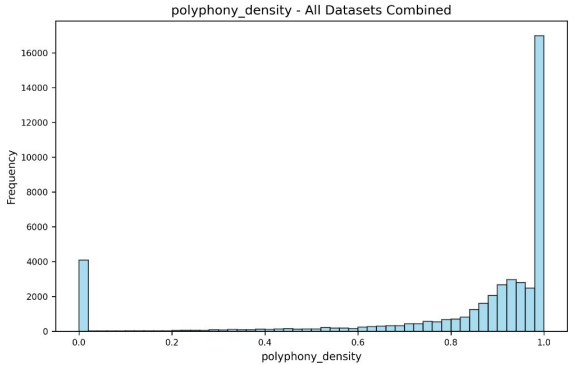


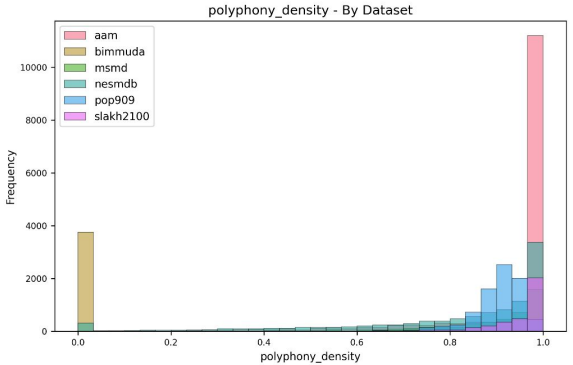
- Rhythmic
  - Hypothesis: Higher variability in note placements (less common rhythms)→ worse transcription accuracy
  - Example: Entropy of Inter-Onset-Intervals
- Harmonic
  - Hypothesis: More complex chord transitions → worse transcription accuracy
  - Example: Average Transition Complexity
- Polyphonic
  - Hypothesis: Higher number of notes played simultaneously → worse transcription accuracy
  - Example: Proportion of piece with 2+ notes being played simultaneously

– Complexity Results

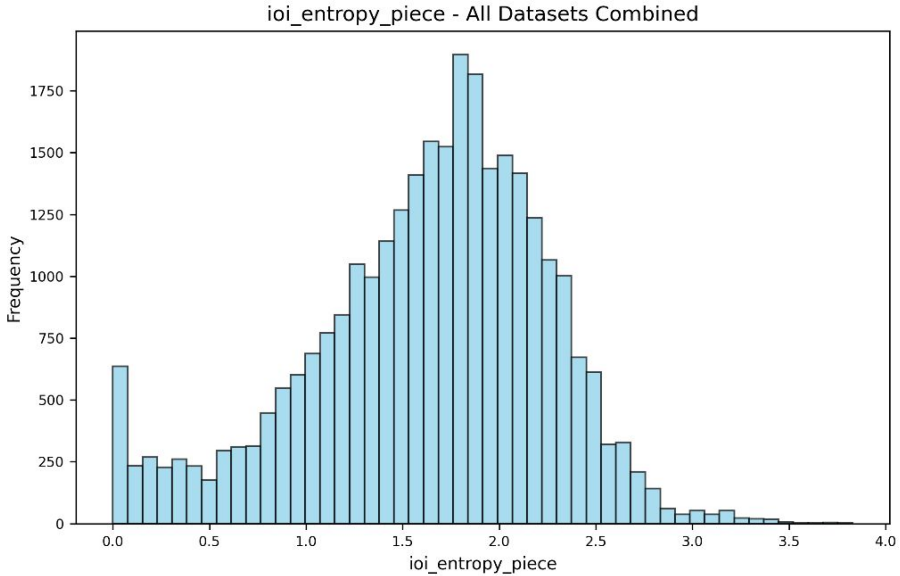


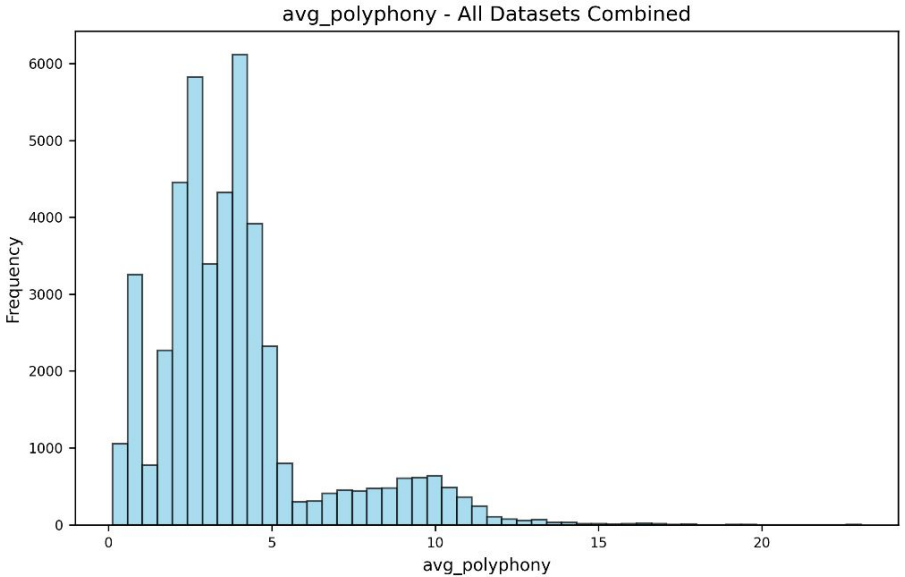




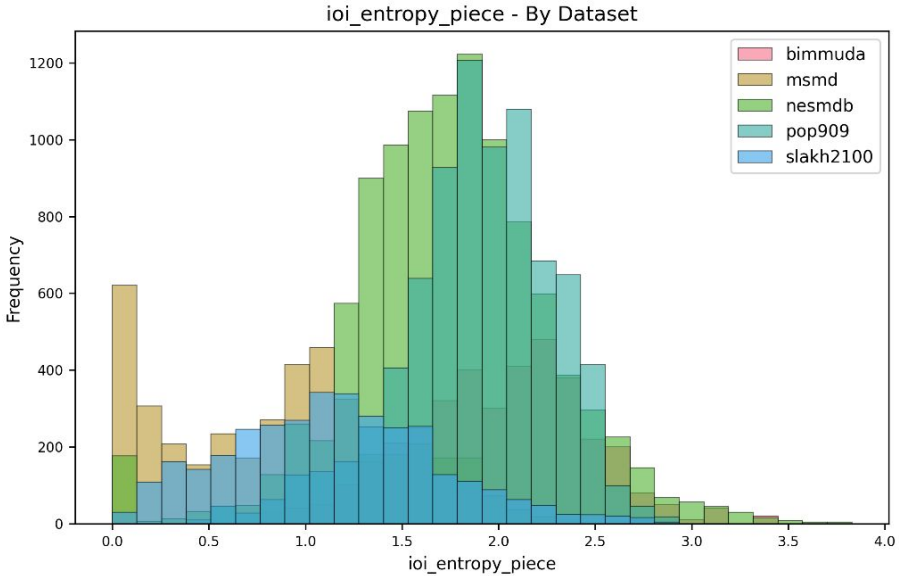


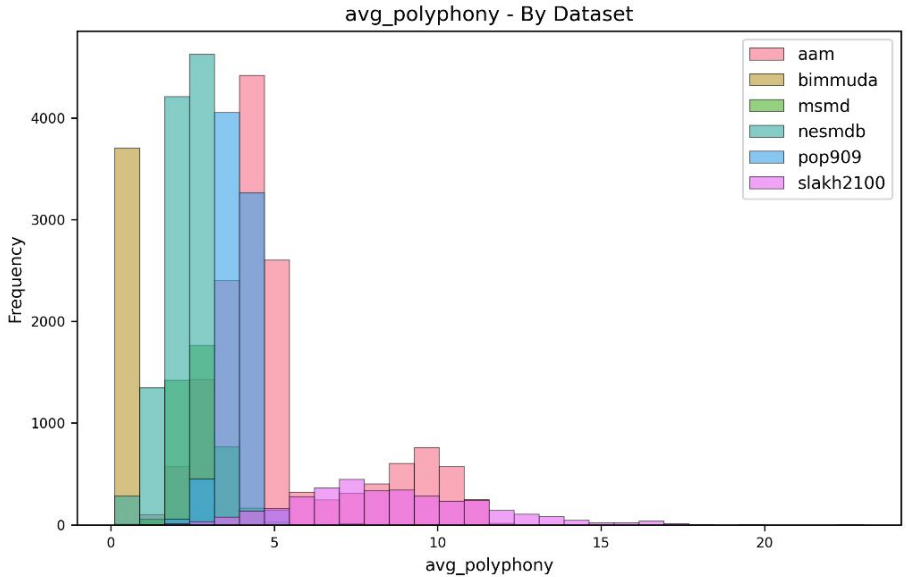
– Complexity Results

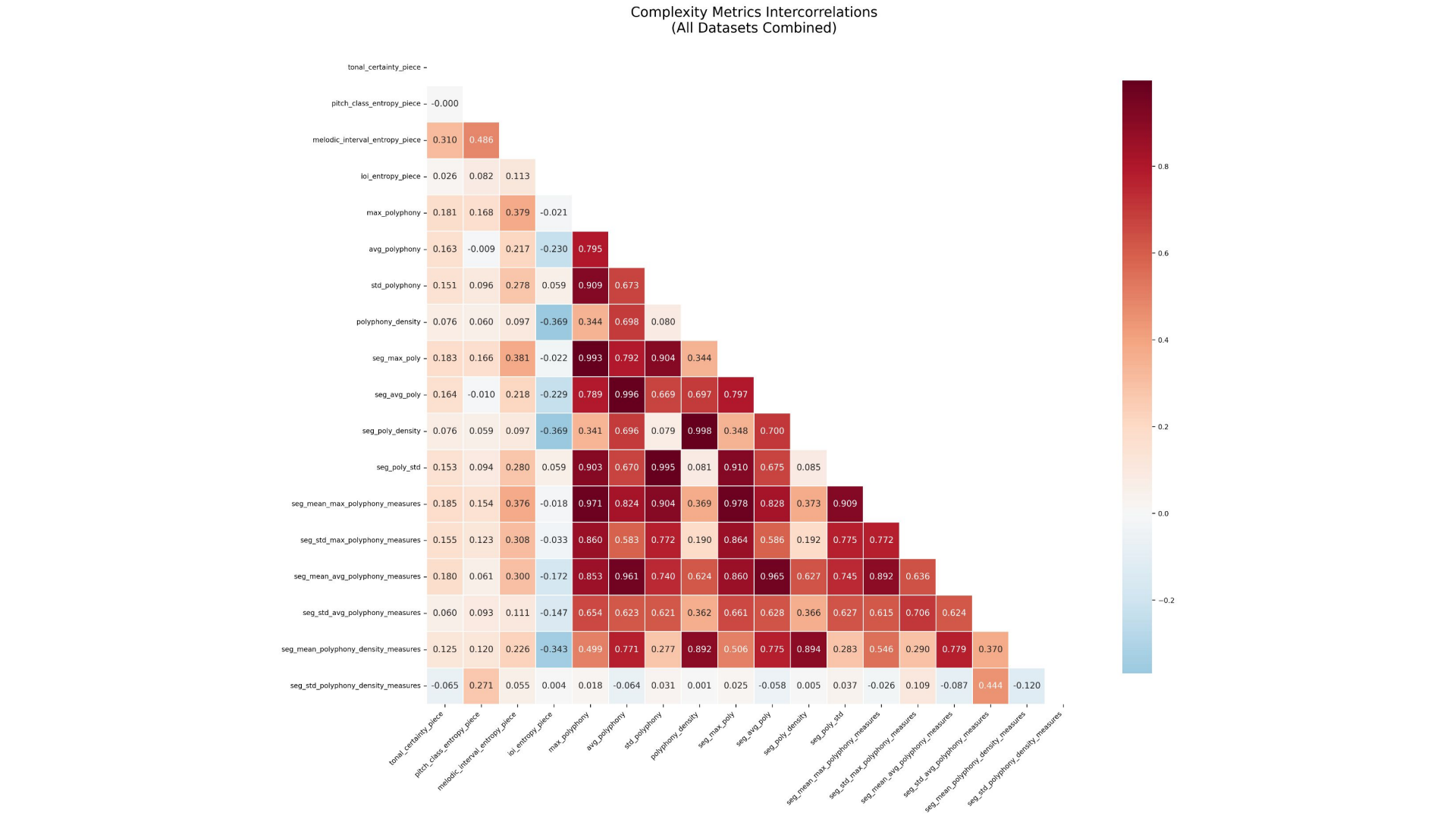




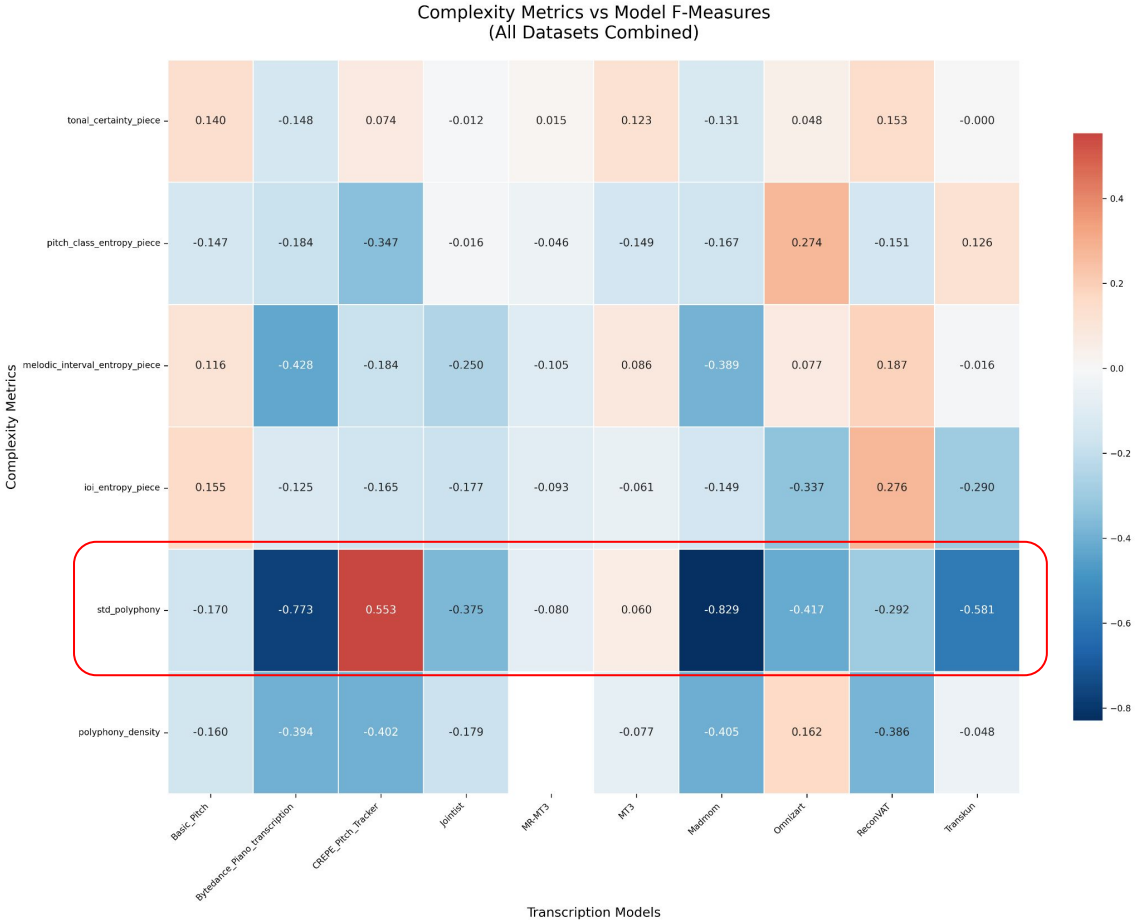
– Complexity Results





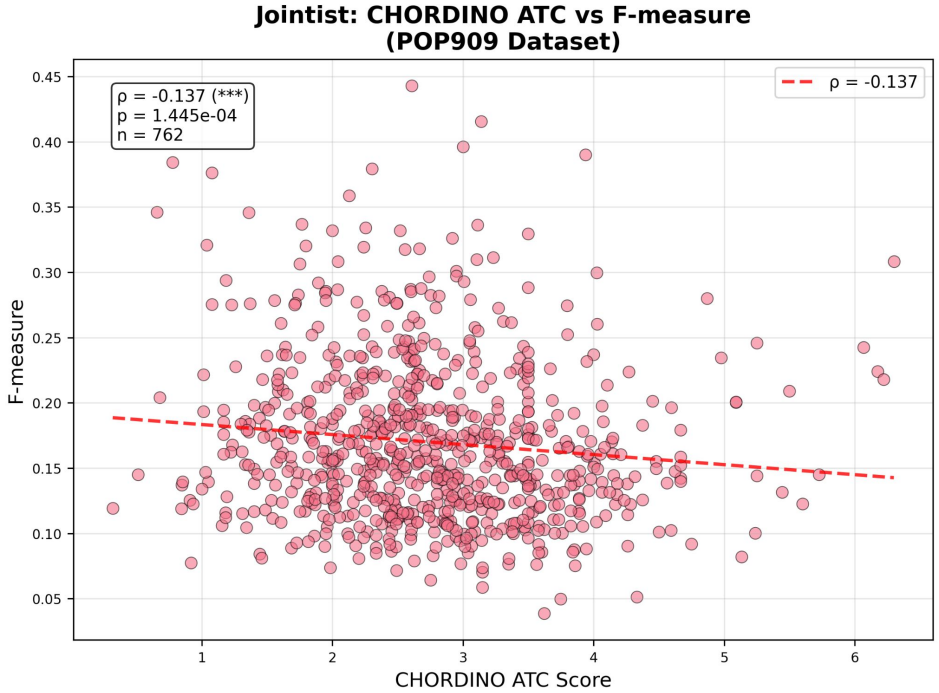

Complexity Metrics Intercorrelations
(All Datasets Combined)
tonal_certainty_piece
pitch_class_entropy_piece -0.000
melodic_interval_entropy_piece 0.310 0.486
ioi_entropy_piece 0.026 0.082 0.113
max_polyphony 0.181 0.168 0.379 -0.021
avg_polyphony 0.163 -0.009 0.217 -0.230 0.795
std_polyphony 0.151 0.096 0.278 0.059 0.909 0.673
polyphony_density 0.076 0.060 0.097 -0.369 0.344 0.698 0.080
seg_max_poly 0.183 0.166 0.381 -0.022 0.993 0.792 0.904 0.344
seg_avg_poly 0.164 -0.010 0.218 -0.229 0.789 0.996 0.669 0.697 0.797
seg_poly_density 0.076 0.059 0.097 -0.369 0.341 0.696 0.079 0.998 0.348 0.700
seg_poly_std 0.153 0.094 0.280 0.059 0.903 0.670 0.995 0.081 0.910 0.675 0.085
seg_mean_max_polyphony_measures 0.185 0.154 0.376 -0.018 0.971 0.824 0.904 0.369 0.978 0.828 0.373 0.909
seg_std_max_polyphony_measures 0.155 0.123 0.308 -0.033 0.860 0.583 0.772 0.190 0.864 0.586 0.192 0.775 0.772
seg_mean_avg_polyphony_measures 0.180 0.061 0.300 -0.172 0.853 0.961 0.740 0.624 0.860 0.965 0.627 0.745 0.892 0.636
seg_std_avg_polyphony_measures 0.060 0.093 0.111 -0.147 0.654 0.623 0.621 0.362 0.661 0.628 0.366 0.627 0.615 0.706 0.624
seg_mean_polyphony_density_measures 0.125 0.120 0.226 -0.343 0.499 0.771 0.277 0.892 0.506 0.775 0.894 0.283 0.546 0.290 0.779 0.370
seg_std_polyphony_density_measures -0.065 0.271 0.055 0.004 0.018 -0.064 0.031 0.001 0.025 -0.058 0.005 0.037 -0.026 0.109 -0.087 0.444 -0.120
tonal_certainty_piece
pitch_class_entropy_piece
melodic_interval_entropy_piece
ioi_entropy_piece
max_polyphony
avg_polyphony
std_polyphony
polyphony_density
seg_max_poly
seg_avg_poly
seg_poly_density
seg_poly_std
seg_mean_max_polyphony_measures
seg_std_max_polyphony_measures
seg_mean_avg_polyphony_measures
seg_std_avg_polyphony_measures
seg_mean_polyphony_density_measures
seg_std_polyphony_density_measures
0.8
0.6
0.4
0.2
0.0
−0.2

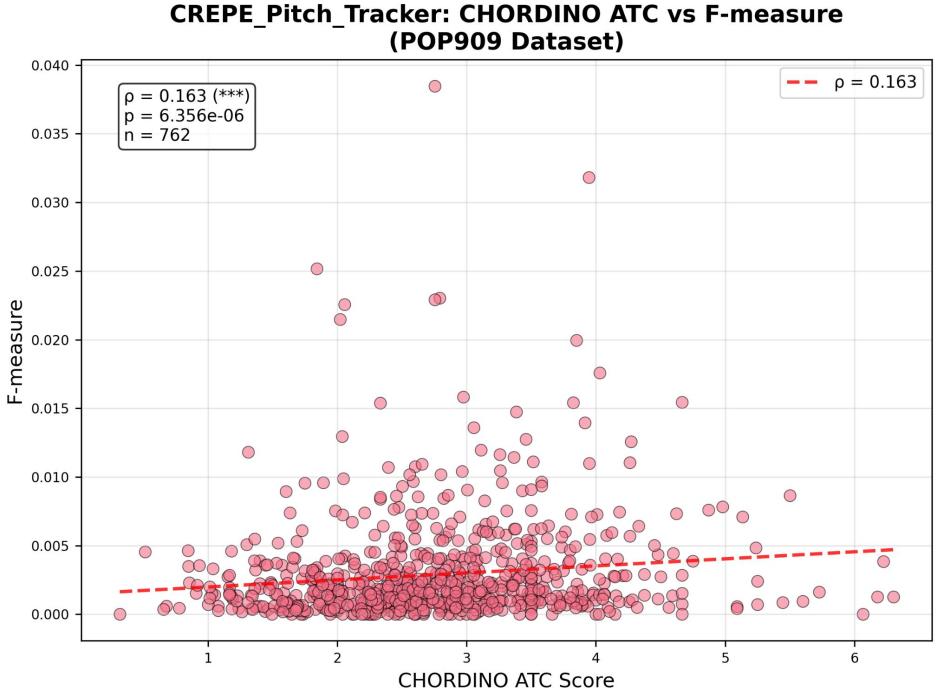
– Complexity Results



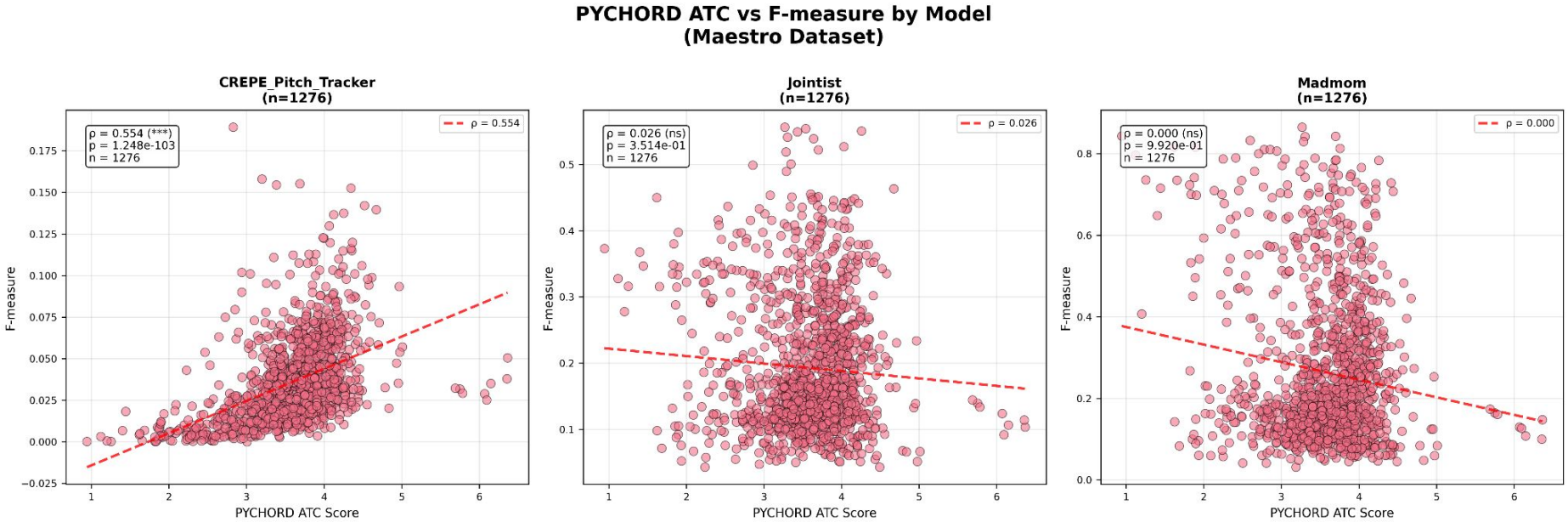
- Positive correlation
  - Higher complexity score -> higher accuracy
- Negative correlation
  - Higher complexity score -> lower accuracy
- Weak correlations
  - Could correspond to the amount of data models were trained on

– Complexity Results

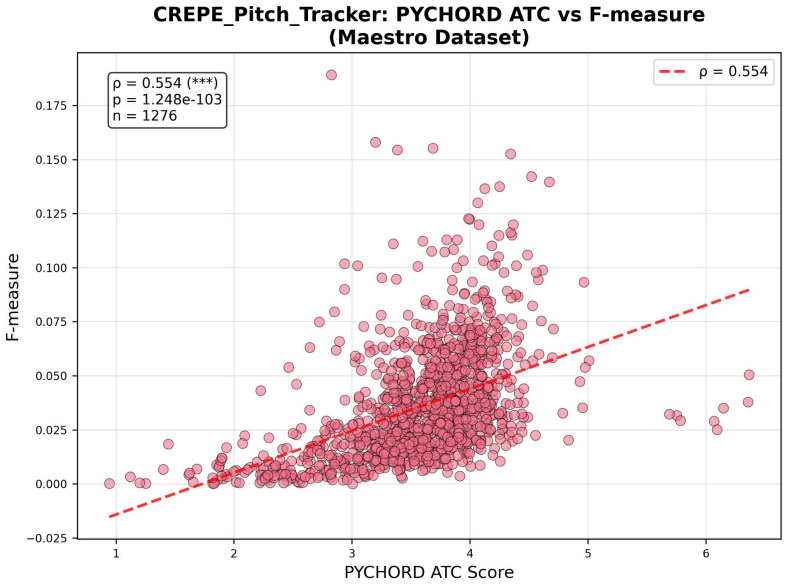


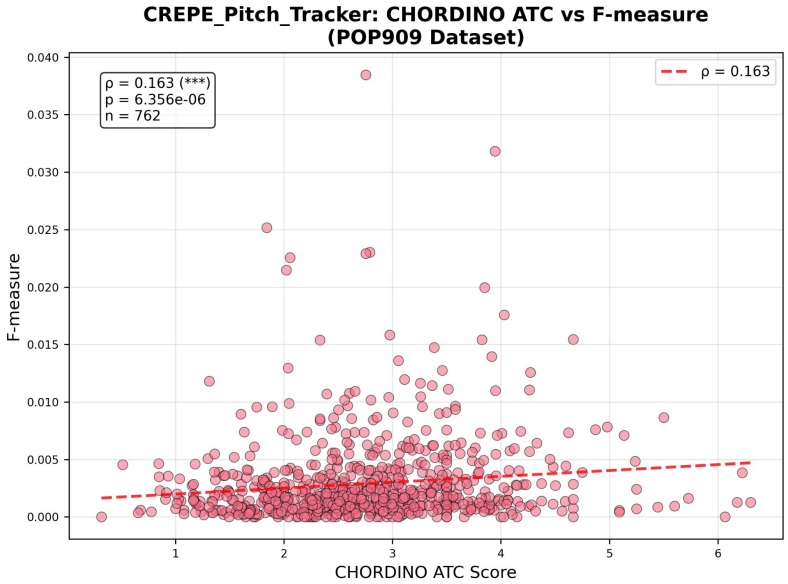


– Complexity Results



– Complexity Results





Results Analysis

- Positive correlation
  - Higher complexity score -> higher accuracy
- Negative correlation
  - Lower complexity score -> higher accuracy
- Weak correlations
  - Could correspond to the amount of data models were trained on

Future Work

NEURAL INFORMATION PROCESSING SYSTEMS

Research Dissemination

The 2025 Automatic Music Transcription Challenge

- Competitors from 5 countries
- Presented results at the AI for Music Workshop at NeurIPS

Paper in progress conducting a literature review in SotA AMT, aiming to publish in IEEE MultiMedia

Drafting a paper to be published to Arxiv

Building an Interpretable Model

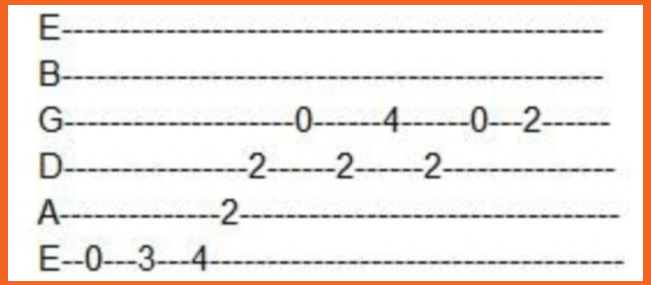
- What is interpretability?
  - Current models are “black boxes”
  - We don’t know what they “learn”
  - Errors are difficult to isolate and fix
  - The overfitting problem

Building an Interpretable Model

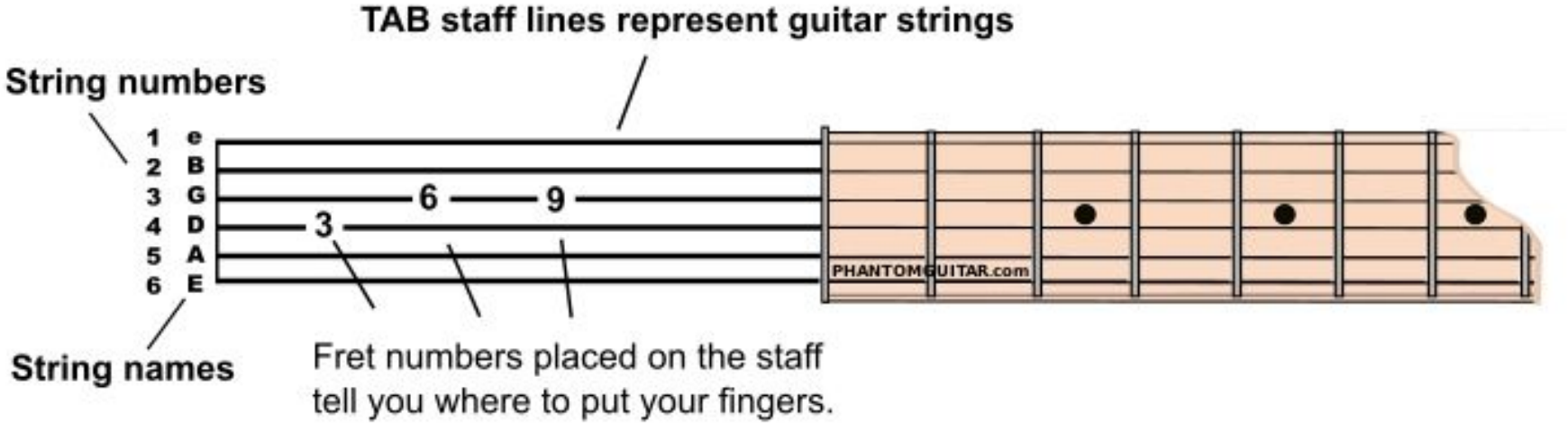
- Our goal is to build an interpretable AMT model:
  - Approaching SoTA accuracy with higher transparency
  - Look inside the model to see why it does what it does

- Planned DDS architecture
  - Differentiable Dictionary Search
  - Deep learning to build a dictionary of possible “notes”
  - Sums and generates the predicted linear sum of the bases

Computer Vision



Converting guitar performance video to tablature



The word **TAB** is placed at the start of a TAB staff so it is not confused with regular music notation..

Source: https://phantomguitar.com/blog/how-to-read-guitar-tab/

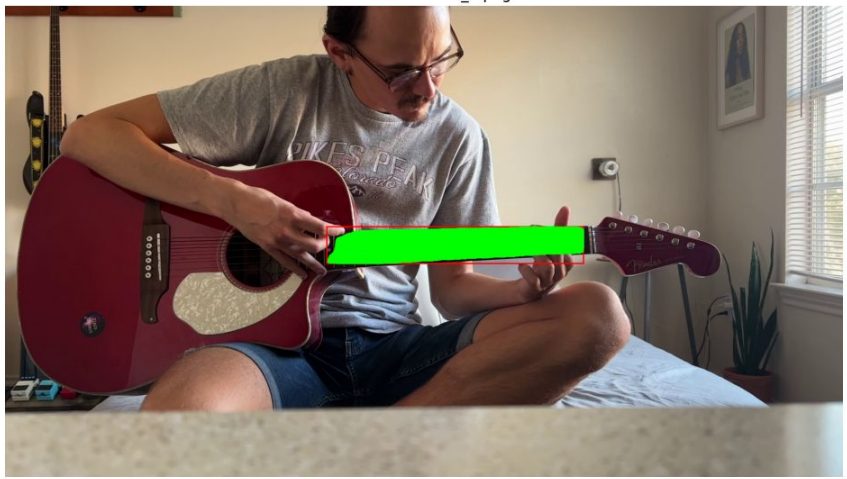
Fingertip Detection (MediaPipe)

31_14.png

34_8.png

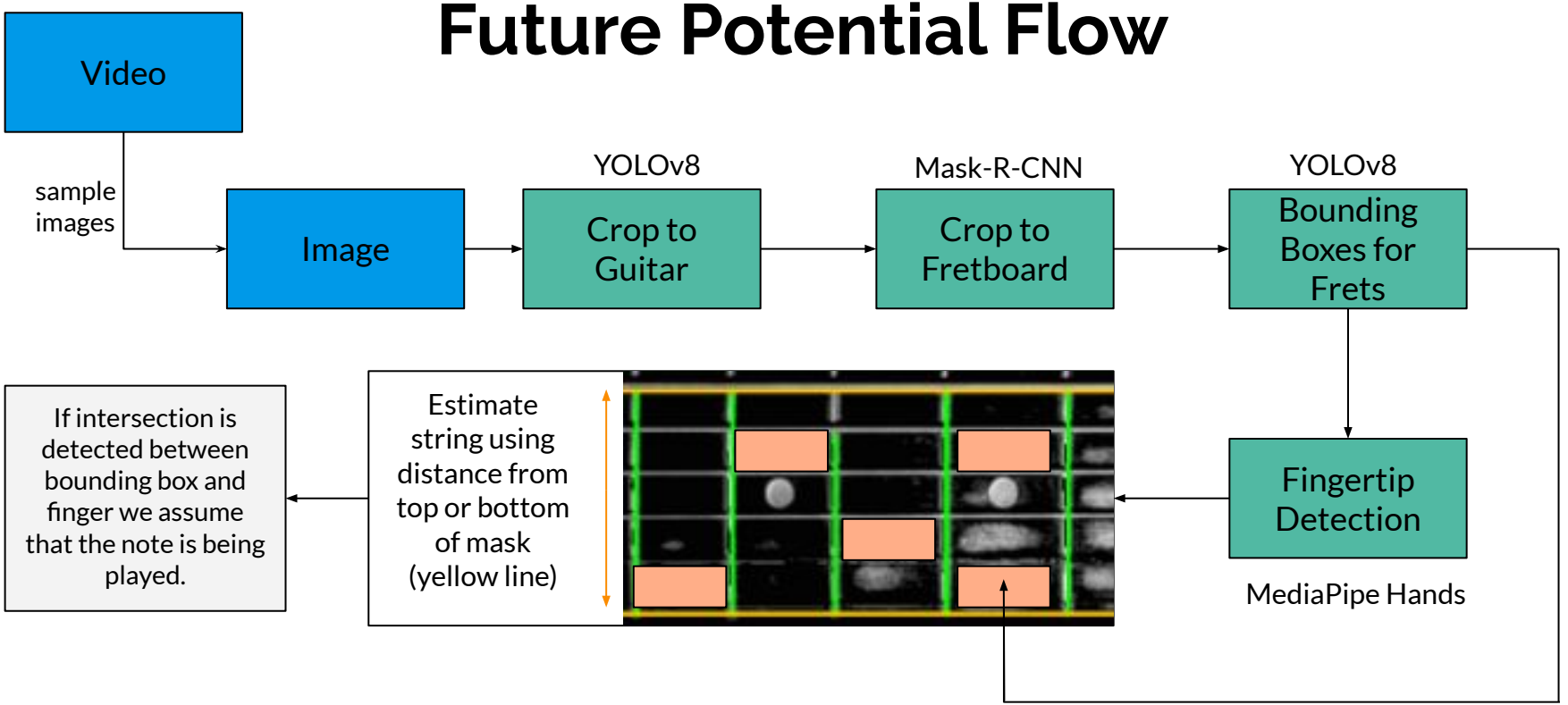
Fretboard Isolation (Mask R-CNN)

Prediction for 31_8.png



Prediction for 34_7.png

Future Potential Flow



Thank you!

Advisors: Professor Yung-Hsiang Lu and Professor Yeon Ji Yun