Understanding Human Music Perception Using Generative AI

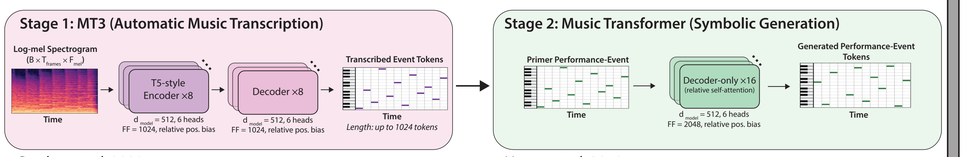
Jamal A. Williams[1], Parisa Vaziri[2], Shih-Lun Wu[1], Richard McWalter[1], Anna Huang[1], Josh McDermott[1,2]

Massachusetts Institute of Technology[1], Harvard University[2]

Introduction

- Human listeners implicitly predict what will happen next in music (predictive coding model of music; Vuust, 2009).
- Musical enjoyment commonly linked to formation of expectations (Huron, 2006).
- Humans are believed to estimate conditional probability of upcoming musical events given previous events.
- Depends on implicit knowledge of culture-specific statistics of music.
- Expectations related to tonality are one well-documented example of musical expectations.
  - Derived through development, enculturation, and musical training (Krumhansl, 2000).
- Existing models capture some patterns of tonal expectation (Pearce, 2005; Temperley, 2008) but are lacking in some important ways:
  - Do not work on raw audio
  - Not powerful enough to estimate the probability distribution of real-world music
- Can generative models that learn musical structure via next-token prediction provide a stronger test of the hypothesis that music cognition arises through continual prediction?

Hybrid Transformer Model



Gardner et al. 2022

- **Training paradigm**: Supervised
- **Training data:** 1,750 hrs multi-instrument datasets of Western music (audio)
- **Objective:** Minimize cross-entropy between true MIDI-event tokens and predicted audio-conditioned token distributions

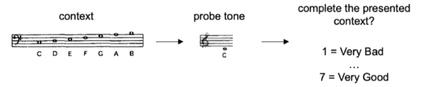
Huang et al. 2018

- **Training paradigm:** Self-supervised
- **Training data:** 10,000 hrs piano recordings (MIDI)
- **Objective:** Minimize cross-entropy between true next token and next-token distributions

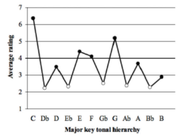
Both pre-trained models were combined to enable experiments using audio stimuli

Probe-Tone Task

Probe-Tone Task

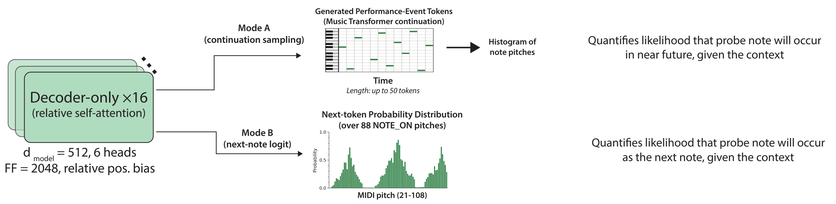


Probe-Tone Profile for C-Major Scale Context

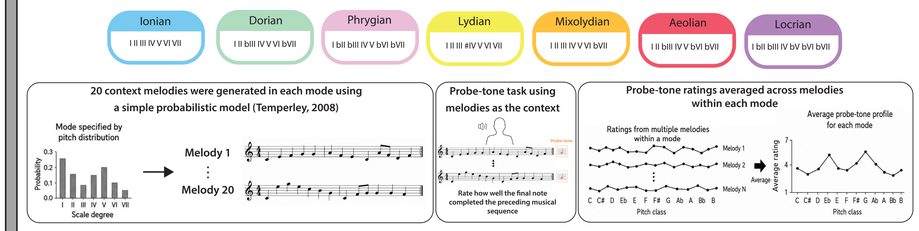


Krumhansl & Shepard, 1979

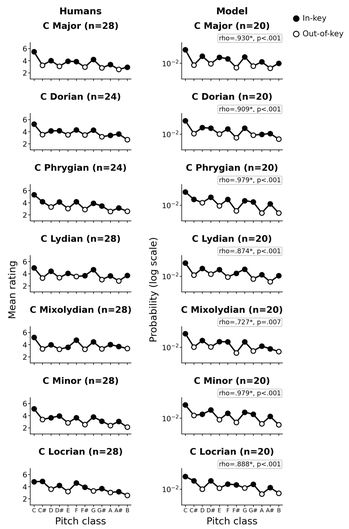
Extracting Model Predictions on Probe-Tone Task



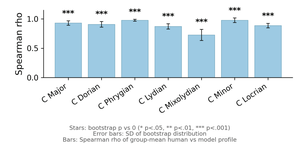
Model Predictions Capture Human Probe-Tone Judgements for Different Musical Modes



Human vs. Model Mode Probe-Tone Results (Next-Note Logits)

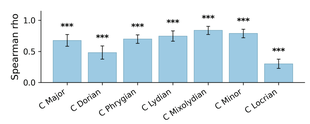


Human vs. Model Mode Probe-Tone Correlation Summary (Next-Note Logits)



Stars: bootstrap p vs.0 (* p<.05, ** p<.01, *** p<.001)
Error bars: SD of bootstrap distribution
Bars: Spearman rho of group-mean human vs model profile

Human vs. Model Mode Probe-Tone Correlation Summary (Continuation Sampling)

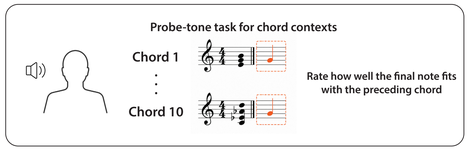


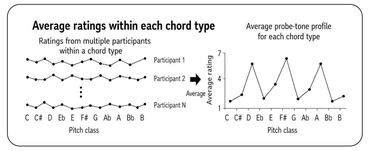
Stars: bootstrap p vs.0 (* p<.05, ** p<.01, *** p<.001)
Error bars: SD of bootstrap distribution
Bars: Spearman rho of group-mean human vs model profile

Main Takeaways

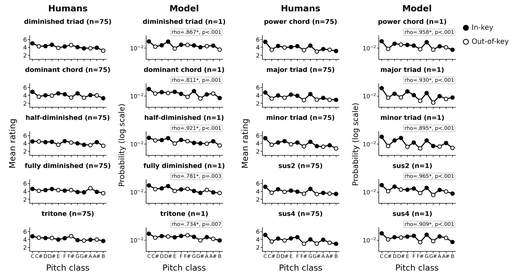
- Human-model profiles both exhibit mode-specific tonal structure
- Human-model ratings for specific pitch classes within modes not always aligned
- Next-note probabilities capture human ratings more consistently than continuation sampling

Model Predictions Capture Human Probe-Tone Judgements for Single-Chord Contexts

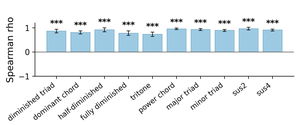




Human vs. Model Single Chord Probe-Tone Results (Next-Note Logits)

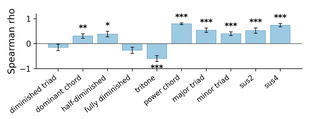


Human vs. Model Chord Probe-Tone Correlation Summary (Next-Note Logits)



Stars: bootstrap p vs.0 (* p<.05, ** p<.01, *** p<.001)
Error bars: SD of bootstrap distribution
Bars: Spearman rho of group-mean human vs model profile

Human vs. Model Chord Probe-Tone Correlation Summary (Continuation Sampling)

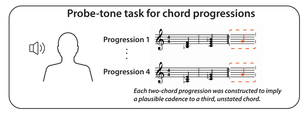


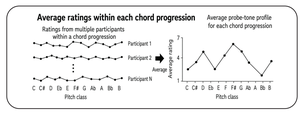
Stars: bootstrap p vs.0 (* p<.05, ** p<.01, *** p<.001)
Error bars: SD of bootstrap distribution
Bars: Spearman rho of group-mean human vs model profile

Main Takeaways

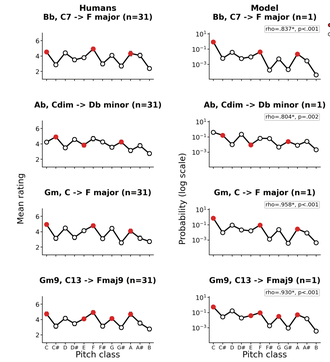
- Considerable human-model correspondence across all chord types
- Human-model similarity is higher for more stable chord types than unstable chord types
- Some differences in details of pitch class ratings between humans and models.
- As with melodic contexts, next-note probabilities capture human ratings more consistently than continuation sampling

Model Predictions Capture Human Probe-Tone Judgements for Chord Progressions

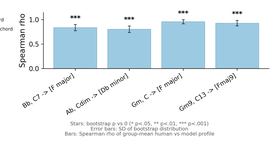




Human vs. Model Chord Progression Probe-Tone Results (Next-Note Logits)

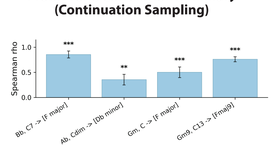


Human vs. Model Chord Progression Probe-Tone Correlation Summary (Next-Note Logits)



Stars: bootstrap p vs.0 (* p<.05, ** p<.01, *** p<.001)
Error bars: SD of bootstrap distribution
Bars: Spearman rho of group-mean human vs model profile

Human vs. Model Chord Progression Probe-Tone Correlation Summary (Continuation Sampling)

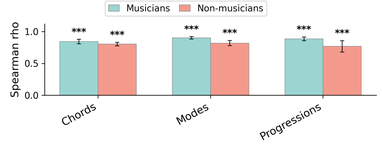


Stars: bootstrap p vs.0 (* p<.05, ** p<.01, *** p<.001)
Error bars: SD of bootstrap distribution
Bars: Spearman rho of group-mean human vs model profile

Main Takeaways

- Model again reproduces much of the structure in human responses, with a few discrepancies
- Pitch classes present in chord context were rated highest for some progressions but not all
- Next-note probabilities capture human ratings for chord progressions more consistently than continuation sampling

Model Predictions Account for Non-Musician Ratings Almost as Well as Musician Ratings



Stars: bootstrap p vs.0 (* p<.05, ** p<.01, *** p<.001)
Error bars: SD of bootstrap distribution
Bars: mean group-mean rho across contexts

Conclusions

- A model trained to predict the next note in Western musical pieces captures human expectation patterns across modes, single chords, and chord progressions.
- Immediate next-note predictions generally correspond more closely to human judgments than note distributions derived from sampled continuations.
- Model-human correspondence is not limited to expert musicians, consistent with idea that most humans implicitly learn to form musical expectations from exposure to music.
- Training a model without assuming MIDI-like perceptual representations might produce better correspondence with human results.
- Future work should explore expectations for rhythm and timbre.

References

1. Vuust, P., Ostergaard, L., Pallesen, K. J., Bailey, C., & Roepstorff, A. (2009). Predictive coding of music--brain responses to rhythmic incongruity. Cortex; a journal devoted to the study of the nervous system and behavior, 45(1), 80–92. https://doi.org/10.1016/j.cortex.2008.05.014
2. Pearce M. T. (2018). Statistical learning and probabilistic prediction in music cognition: mechanisms of stylistic enculturation. Annals of the New York Academy of Sciences, 1423(1), 378–395. Advance online publication. https://doi.org/10.1111/nyas.13654
3. Krumhansl C. L. (2000). Rhythm and pitch in music cognition. Psychological bulletin, 126(1), 159–179. https://doi.org/10.1037/0033-2909.126.1.159
4. Pearce, M. T. (2005). The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition. Doctoral Dissertation, Department of Computer Science, City University of London, UK.
5. Temperley D. (2008). A probabilistic model of melody perception. Cognitive science, 32(2), 418–444. https://doi.org/10.1080/03640210701864089
6. Gardner, J., Simon, I., Manilow, E., Hawthorne, C., & Engel, J. (2022). MT3: Multi-task multitrack music transcription [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2111.03017
7. Krumhansl, C. L., & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. Journal of experimental psychology. Human perception and performance, 5(4), 579–594. https://doi.org/10.1037//0096-1523.5.4.579
8. Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018). Music transformer [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1809.04281

For more information or reprints email jamalw@mit.edu