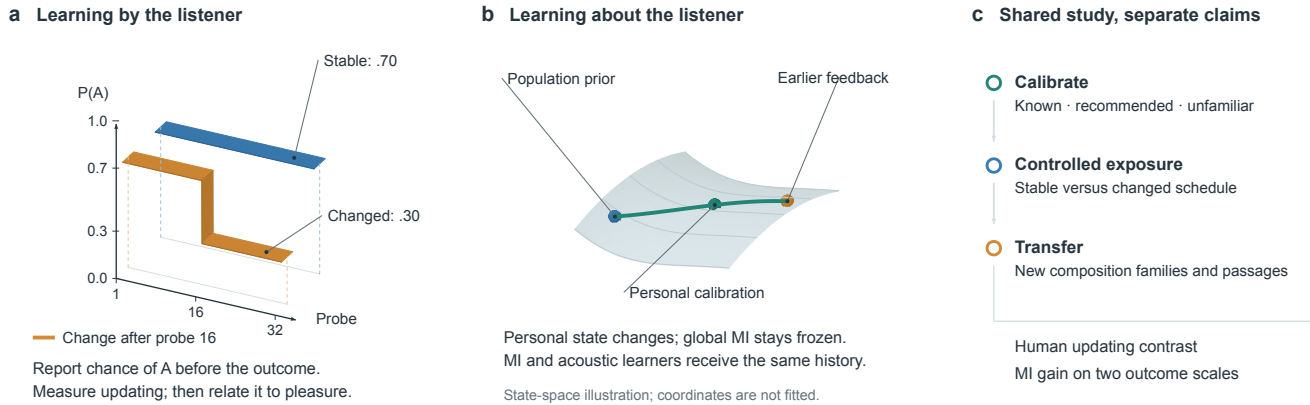


Learning, expectation and musical pleasure: an integrated protocol for evaluating Musical Intelligence

Amaç Erdem | amace@bu.edu | ORCID 0009-0000-1711-0036

Abstract. We propose a study of how listeners learn musical expectations and how those expectations relate to pleasure. The same study will test whether a first version of Musical Intelligence (MI), developed from several strands of the literature, improves prediction of individual responses. Personal calibration draws on familiar music, model recommendations and unfamiliar music, with familiarity and predicted liking recorded separately. Listeners report one response at a time through a mobile interface. Controlled exposure tests changes in expectation; new passages test predictions of continuous pleasure and mean pleasure per excerpt. Acoustic and MI models receive the same listening history, responses and fitting resources. The prespecified criterion is an improvement in correlation greater than 0.10 over the acoustic comparison on each outcome scale. Nine candidate mechanisms will be tested separately and together. A feasibility pilot will determine the measurement procedures and sample size before the confirmatory study. The Supplement gives the detailed methods; a separate language-model analysis benchmark is planned for later work. This protocol reports no participant results.

Author protocol · 12 October 2026 · Supersedes the study and measurement specifications dated 11 October 2026 where listed in Supplement S1. Not a completed preregistration or empirical report.



Pilot qualification → freeze → independent confirmation → revision tested on fresh data. All panels are prospective design schematics.

Fig. 1. Human and model learning. **a**, Assigned continuation probabilities remain stable or change after probe 16. **b**, A geometric illustration separates the population prior, personal calibration and earlier-feedback updates; positions are not fitted states. **c**, Shared observations support independent human and MI claims. All panels describe the proposed design.

1. Scientific rationale and claims

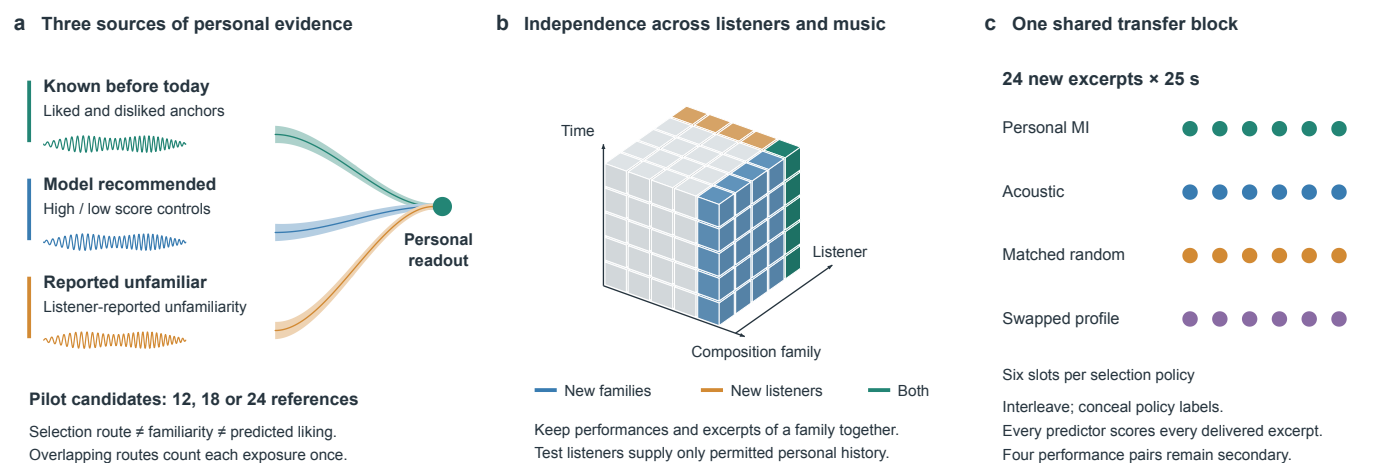
Learning and prediction offer a framework for studying how musical structure and experience shape pleasure[1, 2, 3, 4]. Here we distinguish two questions: how a listener's expectations change with exposure, and how well a model predicts that listener's responses. Evidence of updating can be informative even if MI adds little predictive value. Conversely, accurate prediction alone cannot identify a psychological or neural mechanism.

The human study examines experimentally induced changes in expectation and their relationship to expected and experienced pleasure. The model comparison tests whether MI's fixed transformations add information beyond an equally informed acoustic model. Its nine modules combine ideas from the literature with MI-specific choices of inputs, weights, temporal summaries and

interactions. The source audit documents these choices; it does not estimate a percentage of borrowed formulas or establish their novelty.

There are two confirmatory tests: an exposure-dependent updating contrast and a joint test of MI's predictive contribution across two outcome scales. Secondary analyses examine expectation–pleasure relationships, personalization and individual mechanisms. Estimating an entropy–surprise interaction would require a separate task with fuller probability reports; pleasure ratings alone cannot identify it.

This study evaluates MI v1. Any revisions prompted by its findings will require a new independent evaluation. The Supplement describes the apparatus, analyses and computational sources, preserves the earlier figures and records changes to the protocol.



Candidate counts are starting pilot settings. Cube cells depict grouping, not participant counts or measured values.

Fig. 2. Calibration and independent transfer. **a**, Three routes supply personal evidence; route, familiarity and predicted liking remain distinct. **b**, The index cube illustrates grouping by listener, composition family and time, not sample counts. **c**, Four policies each supply six proposed transfer excerpts; every predictor scores the same delivered targets.

2. Participants, calibration and listening

Cohort. Recruitment is capped at 1,000 distinct eligible registrations, including an initial 60-registration feasibility pilot that will be excluded from confirmation. Adults aged at least 18 will provide study-specific consent and complete delivery and comprehension checks using a compatible phone, browser and headphones[5, 6]. We will record musical training, listening background, recognition, hearing and device information, and missing responses. These background measures will not be used to infer cognitive traits. The initial catalogue consists of rights-cleared classical recordings; extending it will require further sampling and validation. Recruitment arrangements, compensation, institutional responsibility and approvals will be settled before invitations.

Personal calibration. Candidates come from three sources: music known before the study, recommendations based on information collected so far, and music reported as unfamiliar. Known disliked works and high- and low-predicted-preference controls will be included. Selection route, prior familiarity, performance recognition and model score will be recorded separately. The pilot will compare 12-, 18- and 24-reference designs and set the final dose and minimum usable ratings. In each design, seed references precede recommendations; subsequent predictions use only earlier responses. Recommended music may also be unfamiliar, but each exposure is counted once. A preview counts as exposure, not prior familiarity.

Reference ratings initialize a regularized personal readout and are excluded from test outcomes. Selection probabilities, search effort, repetitions and incomplete intake will be retained. Parameters with little personal evidence will be estimated using population priors. Readiness will depend on the calibration coverage rule established in the pilot, replacing the earlier requirement to know exactly 20 works. Reporting musical and demographic coverage will help assess selection bias, which sample size alone cannot remove.

Controlled learning. Listeners first hear two continuation alternatives, A and B, in stable contexts. Before each outcome, a vertical gesture reports the predicted chance of A; the chance of B is its complement. The initial probabilities are (.70,.30). Random assignment either retains this schedule or changes it to (.30,.70), with alternative identity counterbalanced. The start-

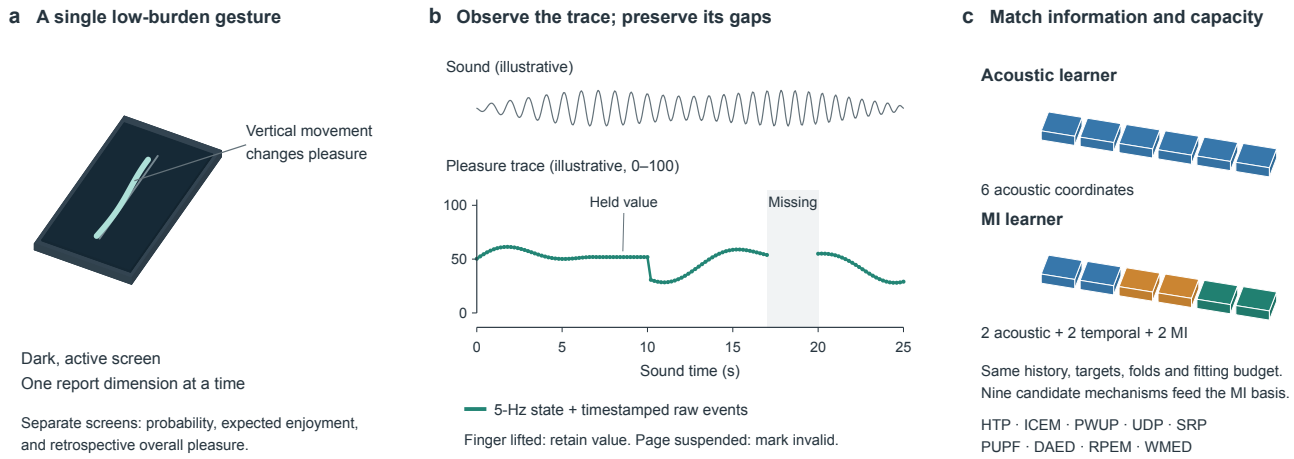
ing pilot schedule contains 16 training exposures and 32 probes, with the change after probe 16. Shorter schedules will be considered against prespecified feasibility and model-recovery criteria. Designated probes collect expected enjoyment and experienced pleasure, one dimension per screen. Probability reports and continuous pleasure reports occur separately.

Transfer. The starting design uses 24 previously unrepresented 25-second excerpts from composition families absent from calibration. Four policies each select six excerpts: personal MI, an equally calibrated acoustic model, matched random selection and MI using another person's matched profile. Every prediction model scores the same delivered excerpts. Policy labels are masked, presentation order is interleaved, and model scores are hidden during listening. Outside-study familiarity is recorded afterwards. Four paired-performance comparisons are secondary. Excerpt counts and duration will be reviewed in the pilot.

Pilot and sample planning

Pilot decisions. The initial 60-registration pilot will assess the gesture interface, three-source calibration, probability comprehension, reporting delay, trial timing, data retention and recovery of the specified effects. Calibration doses and shorter task schedules will be compared on completion, reliability and recovery, rather than the size of an MI advantage. The resulting reference count, usable-reference rule, task counts, response masks, lag treatment, readout basis, regularization and sample size will be recorded in a dated amendment before confirmatory recruitment. An endpoint that cannot be measured or adequately powered will require redesign and another pilot.

The former 55–75-minute estimate and one-point absolute-error target no longer determine primary planning. Session duration will be measured for the mobile design; the 1,000-registration cap remains unless separately amended. Sample-size simulations will consider 300, 500, 700 and 900 analysable participants, accounting for composition-family coverage, temporal dependence, bounded ratings, calibration burden and missingness. The target is adequate error control and at least 90% power at a justified alternative above the 0.10 boundary. Power at the null boundary cannot meet that requirement. The pilot report will state the chosen alternative and achievable precision.



Instrument and signals are analytic design illustrations. A stationary response is valid; missing observations are not filled in.

Fig. 3. Measurement and matched learners. **a**, A proposed mobile gesture changes one report dimension. **b**, An analytic fixture illustrates 5-Hz sampling: a held value remains observed; an interrupted interval is missing. Neither trace is participant data. **c**, Both learners retain six coordinates and identical admissible history; nine mechanisms supply the candidate MI basis.

3. Measurement and model specification

Continuous reporting. During transfer, upward or downward finger movement raises or lowers a 0–100 pleasure value. The active screen may be visually dark; collection will not use an operating-system-locked screen. The pilot will establish gesture sensitivity, anchors, initial-value handling and practice. Value state will be recorded at 5 Hz against delivered audio, together with raw events and validity masks. A stationary response is valid; a suspended page or lost measurement interval is missing. Nominal timestamp resolution does not establish physical timing accuracy[6].

The two primary prediction targets are the aligned within-excerpt pleasure trajectory and its mean over valid, prespecified samples. An immediate overall pleasure rating is secondary and need not equal the trace mean or peak. Recognition and brief unexpectedness reports follow playback. The pilot will estimate reporting delay, repeat reliability, scale use and burden, and compare reporting with passive listening in counterbalanced blocks. A common response-lag rule will apply to all models, selected independently of test correlations. These measurements describe experience during reporting.

Human updating. Let $q_{it}(A)$ be the probability reported before the outcome. For the initially dominant alternative, we compare the mean of probes 25–32 with probes 9–16. The primary effect is the changed-minus-stable difference in this change; a negative value supports updating. The model includes participant and motif-family effects, realized exposure and counter-balanced identity. A shorter pilot-selected schedule will require corresponding non-overlapping windows in the amendment. Expected enjoyment V and experienced pleasure Y are separate reports. Their associations and subsequent changes are secondary, with acoustic, context and exposure controls. The signed discrepancy $Y - V$ is a behavioural measure, not a neural reward-prediction-error measurement[7, 8].

MI v1. The engine, preprocessing, catalogue, module outputs and global learning rules will be fixed before confirmation. The selected modules are HTP, ICEM, PWUP, UDP, SRP, PUPF, DAED, RPEM and WMED[9]. A local audit identified 103 output coordinates across them. These are computational outputs, not 103 independent psychological constructs. In particular, acoustic novelty, spectral entropy and named uncertainty channels do not directly measure a listener’s probability distribution. Prediction of responses to a known recording may use its full audio; claims about anticipation at a given moment require all inputs to be restricted to the past.

Model comparison. Both acoustic and MI representations will be evaluated with population, calibration-only and earlier-feedback-updated readouts. Models receive the same permitted

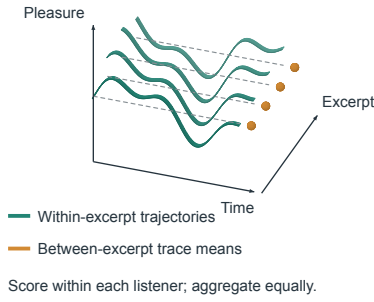
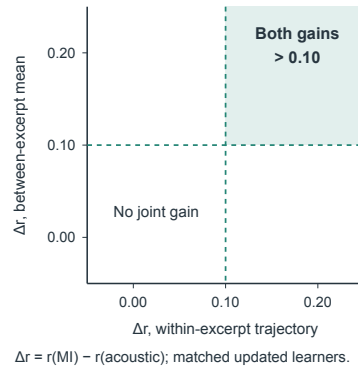
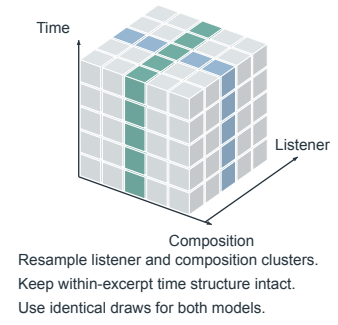
responses, background information, trial order, outer splits and tuning resources. The initial ridge specification has six coordinates: two pilot-fitted components from each acoustic, temporal and mechanism group for MI, compared with six acoustic components, with shrinkage in $\{.1, 1, 10, 100\}$. Temporal and mean readouts use their respective targets. Temporal and nonlinear alternatives matched for capacity will test whether any advantage depends on the readout. Personal-mean, order/repetition and swapped-profile controls assess specificity. All transforms are fitted within the training folds; new test participants contribute only their permitted calibration and earlier responses.

Acquisition and quality

Data collection. Each approved audio representation will be computed once per version. Personal preparation fixes calibration state, selection policy, trial schedule and initial predictions. The phone then presents prepared stimuli and records responses, without live MI, GPU or LLM inference. Records distinguish stimulus family, performance, excerpt and exposure, with audio and software hashes, temporal support, units, consent version and local event order. Audio-context time, browser-monotonic time and server receipt are stored separately so that upload delay is not mistaken for response delay.

Events will be acknowledged only after durable storage. Identical retries return the existing receipt; conflicting duplicates are logged. Bounded local queues retain events awaiting acknowledgement. A replay creates a new exposure and preserves the original response. Visibility loss, audio stalls, device changes and interrupted intervals are recorded. Chronological analysis saves each prediction before revealing its target and updating the model from eligible earlier outcomes. Offline replay will be identified separately from forecasts issued during collection. Saved programs, transforms, masks and folds will allow the same calculation to be rerun.

Quality and missingness. Initial transfer eligibility requires at least 18 of 24 excerpts, each with at least 80% valid samples in the common scoring window; pilot changes will be documented. Constant human responses remain valid for means and error metrics, although within-excerpt Pearson correlation is undefined. Coverage rules are identical across models and depend on the target. A constant prediction for a variable target receives the prespecified zero-correlation score and is counted as a prediction failure. Low pleasure or disagreement with MI is not an exclusion criterion. Reports will include participant flow, endpoint-specific denominators, analyses using all valid observations and missingness by calibration route and background. Primary pleasure outcomes will not be imputed, including after upload failure.

a Two scales of prediction**b A joint practical threshold****c Paired, crossed uncertainty**

Formal confirmation also requires threshold-exceeding evidence with multiplicity control. No empirical correlations or uncertainty intervals are shown.

Fig. 4. Scoring and decisions. **a**, Illustrative trajectories distinguish within-excerpt correlation from correlation across trace means. **b**, The shaded region defines the joint practical target, $\Delta r > 0.10$ on both scales; formal confirmation also requires threshold-exceeding evidence. **c**, Paired crossed resampling preserves time within listener–composition clusters. No performance estimates are shown.

4. Independent evaluation and decisions

Two correlation scales. Pearson correlations between aligned continuous reports and predictions will be calculated for each eligible participant–excerpt pair. Fisher-transformed values will be averaged with equal weight per participant and per eligible excerpt within participant, then back-transformed to obtain r_{time} . Each participant’s correlation across excerpt means will be aggregated with equal participant weight to obtain r_{mean} . The Supplement specifies masks, constant-response cases and minimum coverage. A pooled correlation across all frames or listeners is not the primary analysis.

$$\Delta r_k = r_k(\text{MI}) - r_k(\text{acoustic}), \quad k \in \{\text{time, mean}\}.$$

The primary contrast compares personal MI and acoustic models that both update from earlier feedback. A practically meaningful contribution requires $\Delta r_k > 0.10$ on both scales, supported by the prespecified inferential test. Point estimates and paired intervals will be reported regardless of the decision. An estimate above the threshold without sufficient evidence remains inconclusive; improvement on one scale will be reported at that scale. Mean absolute error, calibration, overall ratings and pleasure under each selection policy are secondary outcomes.

Inference. The analysis will use 5,000 crossed participant-by-composition bootstrap replicates, keeping observations paired across models. One-sided tests against $\Delta r = 0.10$ will use the prespecified centered bootstrap procedure after simulation checks. The larger of the two scale-specific p values tests the joint MI claim. Holm correction at $\alpha = .05$ will cover this claim and the human updating contrast. Two-sided 95% intervals describe uncertainty; confirmatory decisions will account for multiplicity. Secondary mechanism and personalization families will receive separate false-discovery-rate control. Exploratory findings will require later confirmation.

Generalization and model recovery. Primary transfer concerns new composition families in new participants after personal calibration. Prediction of later responses, new families, alternate performances and new people will be reported separately. Outer participant/family splits will be fixed before fitting, with variants, movements, performances and excerpts of a family kept together. Resampling will preserve shared acoustic sources and temporal dependence. Recovery simulations include acoustic-only signals, response-scale differences, smooth autocorrelation, absent or plausible human updating, MI gains below, at and above the threshold, missingness and attrition[10]. These checks will assess identifiability and power at the participant and composition level.

Mechanisms. Acoustic, acoustic-plus-temporal and full selected-mechanism models will be compared with leave-one-module-out variants under matched fitting resources. Related mechanisms will also be tested as groups. Because modules share upstream computations, these analyses estimate contributions of information rather than unique biological causes. Other engine outputs remain exploratory. Weak prediction will prompt checks of delivery, measurement, implementation, target alignment and correspondence with the cited literature before a mechanism is revised.

Interpretation and scope

Interpretation. Human updating and an MI advantage may occur together, separately or neither. The two tests therefore answer different questions. A model failure may reflect its implementation, temporal assumptions, target or precision. It constrains the motivating theory only when the test is sensitive and the implementation represents that theory faithfully. Null findings, disagreements and plausible alternatives will be retained. Changes to weights, inputs or mechanisms define a new MI version and require fresh participants and materials or an untouched reserved evaluation set.

Publication and extensions. This protocol brings the human measurement and analysis methods into one study. The three earlier papers and all 35 original figure sources remain in the editable archive. Supplement S1 records current, pilot-dependent, optional and superseded specifications. The separate documentation/LLM benchmark has its own development and evaluation tasks and primary contrast. Future EEG/MEG/fMRI studies will require independent measurements, synchronization and modality-specific observation models[11]. The mobile study provides no direct measurement of dopamine, neural entrainment or working-memory capacity.

Ethics, access and declarations. Institutional approval, permissions for the recordings, informed consent, withdrawal procedures and compensation will be arranged before recruitment. Contact records will be separated from pseudonymous research data. Permitted materials, executable analyses, synthetic test data and suitably de-identified summaries will be released under stated access conditions[12]. The public draft and interest registration are preparation for a study, not an operating experiment. Amaç Erdem develops MI and proposes this research; affiliation, funding and contributions will be confirmed before submission. AI assistance supported drafting, literature organization and figure/document preparation. No participant dataset, completed operational validation, achieved power or accepted preregistration is reported.

References

- [1] Vuust, P., Heggli, O. A., Friston, K. J. & Kringelbach, M. L. Music in the brain. *Nature Reviews Neuroscience* **23**, 287–305 (2022). doi:10.1038/s41583-022-00578-5.
- [2] Cheung, V. K. M. et al. Uncertainty and surprise jointly predict musical pleasure and amygdala, hippocampus, and auditory cortex activity. *Current Biology* **29**, 4084–4092.e4 (2019). doi:10.1016/j.cub.2019.09.067.
- [3] Gold, B. P. et al. Predictability and uncertainty in the pleasure of music: a reward for learning? *Journal of Neuroscience* **39**, 9397–9409 (2019). doi:10.1523/JNEUROSCI.0428-19.2019.
- [4] Mas-Herrero, E. & Marco-Pallarés, J. Predictive processes shape individual musical preferences. *Proceedings of the National Academy of Sciences USA* **122**, e2500494122 (2025). doi:10.1073/pnas.2500494122.
- [5] Woods, K. J. P., Siegel, M. H., Traer, J. & McDermott, J. H. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics* **79**, 2064–2072 (2017). doi:10.3758/s13414-017-1361-2.
- [6] Bridges, D., Pitiot, A., MacAskill, M. R. & Peirce, J. W. The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ* **8**, e9414 (2020). doi:10.7717/peerj.9414.
- [7] Gold, B. P. et al. Musical reward prediction errors engage the nucleus accumbens and motivate learning. *Proceedings of the National Academy of Sciences USA* **116**, 3310–3315 (2019). doi:10.1073/pnas.1809855116.
- [8] Salimpoor, V. N. et al. Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature Neuroscience* **14**, 257–262 (2011). doi:10.1038/nn.2726.
- [9] Erdem, A. Musical Intelligence: A Glass-Box Architecture for Computational Models of Music Listening. Author manuscript, 8 October 2026. Dated architecture record.
- [10] Wilson, R. C. & Collins, A. G. E. Ten simple rules for the computational modeling of behavioral data. *eLife* **8**, e49547 (2019). doi:10.7554/eLife.49547.
- [11] Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nature Neuroscience* **21**, 1148–1160 (2018). doi:10.1038/s41593-018-0210-5.
- [12] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016). doi:10.1038/sdata.2016.18.