

Relational Bayesian Inference for Musical Pleasure: An Experimental Protocol

Amaç Erdem

Independent researcher

Correspondence: amace@bu.edu · ORCID 0009-0000-1711-0036

Abstract

Can a listener’s response to a question about familiarity or expectation improve the prediction of their subsequent musical pleasure? This protocol proposes a test within Musical Intelligence (MI), using a fixed audio representation and a Bayesian observation model with a persistent listener state. History, direct pleasure reports and auxiliary responses enter separate comparisons, allowing the contribution of relations between tasks to be estimated on the same listening trajectory. Predictions are committed before each listening block and evaluated on later reports. A conditional Gaussian example establishes when an auxiliary observation can reduce predictive uncertainty, while a wrong-sign example shows how an incorrect relation can worsen prediction. The experimental design couples this comparison to a separate, controlled search for associations beyond the dimensions already labelled as pleasure-related. It specifies stimulus provenance, temporal alignment, model controls, held-out source families and simulation-based planning. The personal inference layer and experiment are proposed extensions; the figures show source-audited architecture, analytic calculations and synthetic examples, not participant results.

Study status: Experimental protocol; recruitment, preregistration and ethics review pending.

a Audited audio representation



97 channels in nine R³ groups

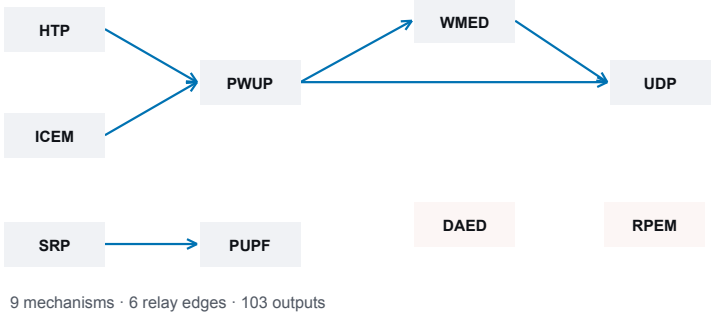
A Consonance · B Energy · C Timbre

D Change · F Pitch/chroma · G Rhythm

H Harmony · J Extended timbre · K Modulation

R³ + eligible T³ arrays supply C³ mechanisms.

b Selected pleasure subgraph



c Proposed personal inference downstream of the fixed engine



F4 memory-labelled scores are audio-derived; a persistent personal state is an extension to be implemented.

Figure 1. Audio representation and the proposed listener model. (a) The nine R³ groups contain 97 channels. Counts describe the inspected implementation. (b) The selected pleasure subgraph has nine mechanisms, 103 outputs and six declared relay edges. DAED and RPEM have no edges within this selected graph; this does not imply isolation in every configuration. (c) History and task responses would update a personal posterior downstream of the fixed engine. The F4 memory-labelled computations currently use audio-derived inputs; persistence across sessions requires an additional observation and storage layer.

1. Introduction

A pleasure rating gives direct evidence about a listener’s response to a particular musical passage. It leaves several explanations open. The passage may be familiar, may fulfil an expectation, or may have an acoustic character the listener tends to enjoy. A second observation can help distinguish these explanations even when it does not ask about pleasure. The question is whether the distinction improves a prediction that has not yet been scored.

This protocol examines that question using Musical Intelligence (MI). MI supplies named acoustic coordinates, temporal operations and listening-related calculations with inspectable dependencies. The proposed extension adds a personal state inferred from a listener’s history and responses. Its central test concerns information transfer: does a familiarity report or an expectancy response improve the forecast of later pleasure beyond the information already supplied by previous pleasure ratings?

Joint Bayesian models and multi-task prediction already provide ways to combine observations through shared latent variables or covariance structure (Turner et al., 2013; Bonilla et al., 2007). The proposed contribution is a specific use of these principles in

an inspectable music-listening architecture, together with an experiment that can reject the assumed relations. A larger collection of scores would not by itself establish a contribution. The relations must improve a defined prediction under controls for history, direct feedback, model capacity and stimulus identity.

Musical familiarity and expectation are plausible starting points because each has an established relationship with pleasure, although neither supplies a universal account of it (van den Bosch et al., 2013; Cheung et al., 2024). Specific musical anhedonia further cautions against assuming that intact perceptual processing necessarily produces reward (Mas-Herrero et al., 2014). These observations motivate a small initial model whose relations can vary across listeners. They do not justify treating every mechanism name in MI as a measured psychological variable.

A second aim is exploratory. An acoustic or computational dimension with no prior pleasure label may carry useful information about a listener’s response. Such a result could reveal a missed predictor, an interaction with context, or a limitation of the current model. The protocol therefore separates a confirmatory test

of specified relations from a broader search for additional associations. The latter has its own selection, multiplicity and replication requirements.

The study concerns inference about a listener. Changes in the model's posterior do not establish that the listener has learned. Likewise, the fixed audio representation does not make the entire procedure external to statistical learning: Bayesian updating is often included under that broad term. The precise commitment is that this experiment uses no deep neural network and does not retrain an audio representation from participant ratings. It tests whether explicit probabilistic relations between observations are useful.

2. Representation and observation model

Implemented computations and proposed state

The inspected R³ implementation contains 97 channels in nine groups. T³ supplies requested temporal operations, and C³ combines these quantities into named calculations. The selected pleasure-related subgraph contains nine mechanisms and 103 outputs, connected by six relay edges (Figure 1). The broader C³ inventory contains 84 mechanisms and 889 output dimensions. These numbers describe the repertoire and its dependencies; they do not count independent psychological processes or independent statistical tests.

The distinction matters especially for memory. In the inspected HCMC implementation, eleven bounded outputs are computed from acoustic, temporal and relay inputs. The interface does not accept a participant response or a persistent personal history. A memory-related name therefore cannot establish that a score stores what a particular listener has heard. The proposed study would add that missing information explicitly, with a separate data model and update rule.

For recording s and time t , let the fixed engine produce

$$\mathbf{z}_{st} = F_{\text{MI}}(a_s)_t, \quad (1)$$

where a_s is the verified waveform. A declared subset of these coordinates and temporal summaries enters the observation model. The engine configuration, normalization and temporal support are frozen before the confirmatory evaluation. Full-recording audio may be available for an offline listening-response prediction; claims about online audio forecasting would require an additional restriction on every upstream operation.

Let θ_u denote listener u 's state and $\mathbf{h}_u$ their recorded history profile. A useful starting specification is

$$\theta_u \sim \mathcal{N}(\mathbf{B}\mathbf{h}_u, \Sigma), \quad (2)$$

with a population prior when history is unavailable. The initial state should be small enough to recover in simulation. Candidate components include an individual hedonic intercept, familiarity and expectancy sensitivities, and a limited set of weights on selected MI coordinates. Their number, priors and permissible covariance structure must be fixed after development and recovery checks, before confirmatory data are collected.

Separate likelihoods for separate responses

A response from task k is an observation of that task, with its own measurement scale:

$$r_{uk} \sim \mathcal{L}_k(h_k(\mathbf{z}, \theta_u, \mathbf{c})). \quad (3)$$

Here $\mathbf{c}$ records context, such as block, presentation order and playback conditions. The proposed familiarity scale has five ordered categories. An ordered-logit likelihood respects their ordering without assuming equal distances. The expectancy task asks the listener to choose between two continuations and uses a Bernoulli likelihood, or an equivalent two-option softmax. The primary pleasure response is continuous on a 0–100 scale. A censored Gaussian observation model with centre $100 \logit^{-1}(\eta_P)$ is a candidate specification, including explicit probability at the scale boundaries.

The task links need not all depend on every state component. Their shared dependence determines what one task can reveal about another. A pleasure intercept can remain distinct from perceptual sensitivity, and a covariance may be positive, negative or absent. Fixing all such relations to a positive sign would rule out potentially informative dissociations before the study begins. Recognition accuracy, reported surprise and urge to move could be studied later, but they are not interchangeable with either familiarity or felt pleasure.

Given the observations D available at an update, the posterior is

$$p(\theta_u | D, r_{uk}) \propto p(r_{uk} | \theta_u, \mathbf{z}, \mathbf{c}) p(\theta_u | D). \quad (4)$$

Predictions integrate over this uncertainty rather than substituting a single fitted state. Shared coefficients are estimated only from development data. If the primary implementation fixes them for computational reasons, sensitivity analyses must propagate uncertainty using frozen draws from their development posterior. Both the conditional approximation and its consequences for calibration should be reported.

The primary model treats the relevant listener state as stable within a session. This is a testable simplification. A dynamic-state extension could distinguish persistent preference from fatigue or changing attention, but it also introduces parameters that may be hard to identify. It should be evaluated separately, after demonstrating that the proposed tasks can recover the simpler state.

3. When an auxiliary response is informative

A conditional Gaussian example

Consider future pleasure P and an auxiliary response Z , after conditioning on the information already in D . Suppose their joint distribution is Gaussian with conditional means μ_P, μ_Z , variances v_P, v_Z and covariance c_{PZ} . Observation noise is included in v_Z . For $v_Z > 0$,

$$\mathbb{E}[P | D, Z] = \mu_P + \frac{c_{PZ}}{v_Z}(Z - \mu_Z), \quad (5)$$

$$\text{Var}(P | D, Z) = v_P - \frac{c_{PZ}^2}{v_Z}. \quad (6)$$

The second expression gives an exact reduction under the stated model (Rasmussen and Williams, 2006). A response carries no additional information in this Gaussian example when its conditional covariance with future pleasure is zero. A negative covariance can be as informative as a positive one. The relevant covariance is conditional on previous observations; a large raw correlation may contribute little once history and past pleasure are known.

For several auxiliary responses, write their conditional covariance matrix as V and their covariance with P as $\mathbf{c}$. If V is invertible,

$$\text{Var}(P | D, \mathbf{Z}) = v_P - \mathbf{c}^T V^{-1} \mathbf{c}. \quad (7)$$

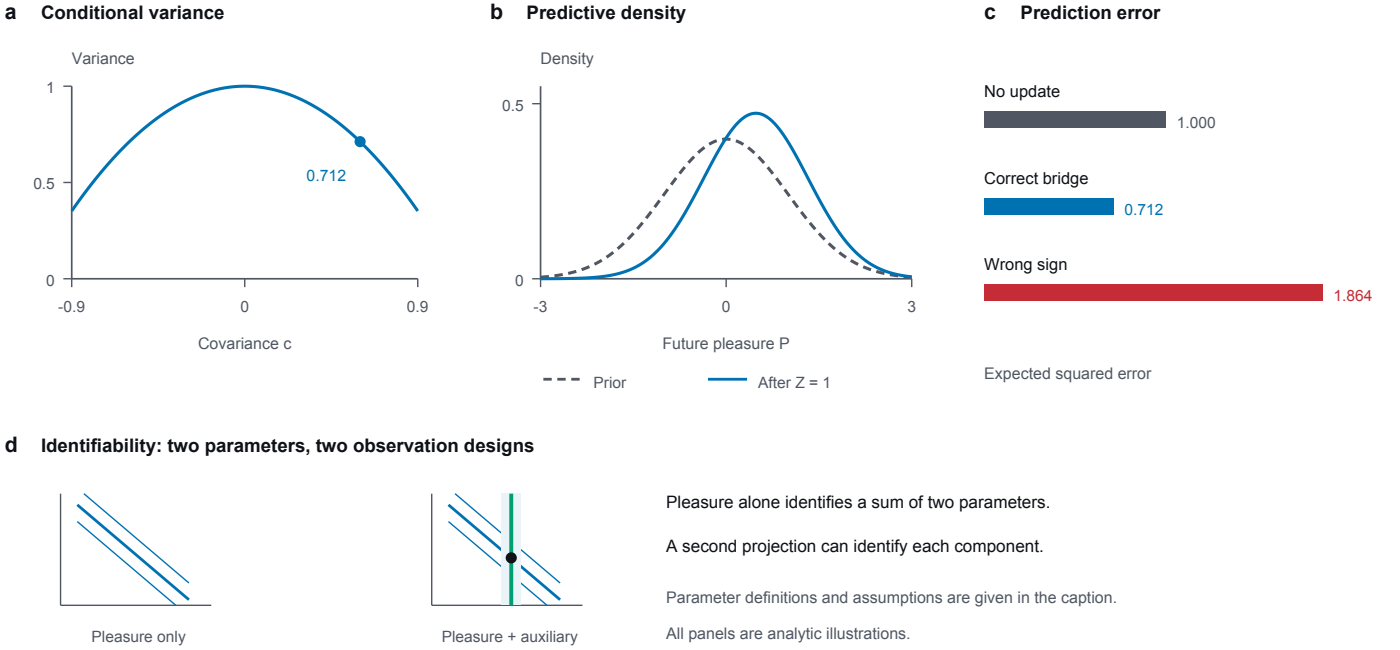


Figure 2. Information transfer under a specified model. (a–c) The illustrative conditional Gaussian has $v_P = 1$, $v_Z = 1.25$, $c_{PZ} = 0.60$ and zero prior means. Its update gain is 0.48 and its posterior variance is 0.712. The density in (b) conditions on $Z = 1$. The expected squared errors in (c) are 1 without updating, 0.712 with the correct relation and 1.864 when the update uses the opposite sign. These are analytic expectations, not experimental estimates. (d) Observing $P = \alpha + \beta$ alone leaves a line of parameter solutions; an additional observation of $Z = \alpha$ supplies a second projection. The sketches depict identifiability, not sampled posteriors.

This expression accounts for dependence among the responses. Repeating an almost identical measurement does not provide the information that would be obtained from several independent measurements. In practice, near-singular covariance estimates also require regularization and sensitivity analysis; adding tasks can otherwise make estimation less reliable.

Figure 2 uses $v_P = 1$, $v_Z = 1.25$ and $c_{PZ} = 0.60$. The gain is 0.48 and the conditional variance is 0.712. To see the cost of a wrong relation, let a linear forecast use gain a on a centred Z . Its expected squared error is

$$R(a) = v_P - 2ac_{PZ} + a^2v_Z. \quad (8)$$

The correct gain minimizes this risk. Reversing its sign gives $v_P + 3c_{PZ}^2/v_Z = 1.864$, which is worse than ignoring Z . Thus, a Bayesian update is not an empirical guarantee of improvement. The relation encoded by the model must be adequate for the observations to which it is applied.

Expected information and realized outcomes

The general information statement is an expectation over possible observations:

$$I(P; Z | D) = \mathbb{E}_{Z|D}[\text{KL}(p(P | D, Z) \| p(P | D))] \geq 0. \quad (9)$$

For squared-error prediction under a coherent model, the corresponding expected Bayes-risk reduction follows from total variance:

$$\begin{aligned} \text{Var}(P | D) - \mathbb{E}_{Z|D}[\text{Var}(P | D, Z)] \\ = \text{Var}_{Z|D}(\mathbb{E}[P | D, Z]) \geq 0. \end{aligned} \quad (10)$$

These statements do not say that every realized response narrows every credible interval. Outside the fixed-covariance Gaussian case, a surprising response can expose uncertainty and increase the realized posterior variance. Nor do the statements ensure accuracy when the likelihood, shared covariance or measurement process is wrong. Held-out predictive performance remains necessary.

An auxiliary task can also resolve parameter ambiguity. If pleasure depends only on $\alpha + \beta$, pleasure observations cannot separate the two components. A second task sensitive to α adds an independent projection and can make both identifiable (Figure 2d). An informative prior may stabilize the first model, but it does not supply a second data constraint. Recovery simulations should determine which distinctions the actual task design can support (Wilson and Collins, 2019).

Derived scores are not additional observations

Suppose $Q = g(X)$ is a deterministic transformation of an already observed quantity. Then

$$I(P; Q | X, D) = 0. \quad (11)$$

A derived MI score may express a useful hypothesis or make a limited model easier to fit. It still cannot create information beyond its complete inputs. Treating several algebraically related scores as independent evidence would exaggerate precision. The proposed advantage of auxiliary tasks is different: they introduce observations of the listener that are absent from the waveform and from previous pleasure reports.

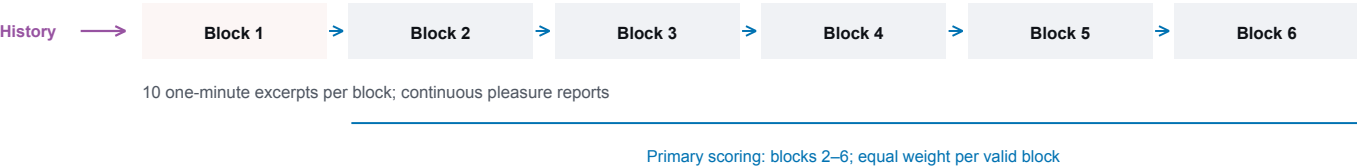
This distinction also limits what a large architecture can establish. A useful composite may be easier to interpret than its acoustic constituents, or impose a helpful constraint on a small dataset. Whether it adds predictive value must be tested against those constituents under the same protocol. Interpretation, convenience and incremental information are separate empirical questions.

4. Experimental design

Recording identity and listening history

Spotify and Apple Music records would identify titles, artists and a participant’s reported or exported listening history. Audio would be obtained separately from a source that permits the intended research use. A title match is insufficient: covers, remasters, edits and live versions can have different waveforms. The matching record should include artist, version, duration, an ISRC where available and an audition check. Unresolved matches are excluded or explicitly marked as a different recording.

a Within-listener chronology



Commit forecasts before each block. Reports from that block enter only the next update.

b Evidence available to each virtual model

Model	History	Pleasure	Auxiliary	Bridge	Comparison
A Population	—	—	—	—	Baseline
B History	Yes	—	—	Off	B versus A
C Direct feedback	Yes	Past	—	Off	C versus B
Separate-task control	Yes	Past	Past	Off	Diagnostic
R Relational	Yes	Past	Past	On	R versus C
R-yoked	Yes	Past	Yoked	On	Control

Figure 3. Block chronology and model comparisons. (a) All models forecast the same six listening blocks. Predictions for a block are committed before its first excerpt; responses from that block become available only at the following update. Primary evaluation uses blocks 2–6. Familiarity prompts and continuation probes occur at prespecified positions. (b) The comparison separates population information (A), listening history (B), direct pleasure feedback (C), separate task models ($C^\perp$), relational updating (R) and yoked auxiliary responses. The primary contrast is R versus C; equal task exposure keeps participant burden the same across these virtual models.

Each waveform receives a hash, a parent recording identity and a source-family identifier. Excerpt boundaries, transformations, loudness processing, playback settings and MI configuration are recorded alongside hashes of the resulting arrays. Alternate excerpts and transformed versions of the same source remain in one family for partitioning. Otherwise, an apparently new test track may reproduce material already available during development.

History is voluntary and is limited to information needed for the study. The record should preserve its source, cutoff date, missingness and uncertainty. Playlist membership alone does not establish listening, liking or an autobiographical memory. Music-evoked memory is a plausible dimension of experience (Janata, 2009), but inferring personal events from a song list would exceed these observations. Private narrative content is unnecessary for the primary experiment.

The stimulus panel combines participant-selected familiar music with source families sampled independently of model performance. Novelty is checked using the participant’s report rather than inferred from catalogue metadata. Selection criteria, balancing variables and the treatment of uncertain familiarity are fixed before confirmatory collection. Both the history model and its comparators predict the same panel, so preferential selection for one model cannot create its apparent advantage.

Session and task sequence

The provisional session contains sixty one-minute scored excerpts in six blocks of ten (Figure 3). Participants report felt pleasure continuously on a 0–100 scale. Two familiarity prompts and two continuation probes follow selected excerpts in each block. Probe audio is unscored and is distinguished from the sixty minutes of primary listening. Setup, practice and breaks add to this duration; an initial allowance of approximately eight minutes for probes must be checked in the pilot rather than treated as an established burden estimate.

The familiarity prompt asks whether the excerpt was known before the current session. The continuation task presents two alternatives under a fixed presentation rule. Instructions, timing and response windows are piloted for comprehension and then frozen. Prompt locations should avoid systematically coupling a

task to only one position, intensity range or music category. The sequence and any counterbalancing are generated before predictions are evaluated.

Eligible adults would be recruited under a prespecified hearing and language protocol. Musical background, age range and recruitment source must be recorded, with any quotas defined before collection. Hearing criteria, exclusion rules and the minimum usable response coverage are unresolved design choices to settle in the pilot. The main study should begin only after the practical burden, recovery performance and sample-size justification support the resulting specification.

Practice includes scale use and examples that are excluded from evaluation. Participants are told to report pleasure as experienced, including low or neutral responses; continuous movement is not required. A stable slider position can therefore be valid. An untouched default before the first deliberate response, however, is not automatically a pleasure observation. The interface must record enough events to distinguish these cases.

Evidence available to each model

Model A uses a population readout. Model B adds the history-informed prior and remains fixed during the session. Model C starts from the same history prior and updates from completed pleasure reports. Model R receives the same history and pleasure information as C, plus completed auxiliary responses linked through the specified listener state. The primary contrast, R versus C, estimates the value of this relational use of auxiliary observations.

A separate-task model, $C^\perp$, receives the same task responses but has no cross-task connection into the pleasure state. It checks that access to an additional task file does not itself alter predictions through an unintended implementation route. A yoked control supplies R with another listener’s auxiliary responses while preserving task and block structure. The assignment must respect the conditions needed for exchangeability, including any stimulus-specific response dependence; a naive shuffle across unrelated probes would change more than personal correspondence.

Before block 1, B, C, $C^\perp$ and R must have the same history-conditioned marginal pleasure prediction. Their later differences

can then be attributed to the observations made available by the protocol, subject to the declared model differences. Common parameters, priors and fitting budgets are matched wherever possible. Remaining capacity differences are reported and assessed with a generic joint-model comparator. Additional parameters should not be credited as evidence for a particular MI mechanism.

These are virtual models evaluated along one participant trajectory. The participant performs one set of tasks; each model receives a defined subset of the resulting evidence. This design holds music, reports and burden constant for the primary comparison. It estimates how an inference procedure uses observations under this panel. It does not estimate what would happen if the participant had never been asked the auxiliary questions.

Commitment and update records

Before each block, the system saves all predicted trajectories and their predictive distributions, model identity, state version, available-data cutoff and code version. The record is immutable for scoring purposes. The next update consumes only observations from completed blocks. A current-block response cannot alter a prediction already attributed to that block, even when the interface displays a live estimate for demonstration.

Raw audio time and raw response events are retained separately. The audit log records pauses, hidden-window intervals, interruptions, device changes and probe onset. A replay should reproduce both the predictions and the exact set of observations available when they were committed. If a software fault breaks this ordering, the affected block is handled according to a prespecified rule shared by all models.

Development sequence

Development proceeds from an executable observation model to a replayable session, followed by recovery simulation and a feasibility pilot. Each stage should produce an identifiable configuration rather than silently replace the preceding one. The shared engine can remain fixed while candidate likelihoods and personal-state dimensions are compared on development data. Once selected, their formulas, priors and update tolerances are frozen together.

The replay test uses a synthetic participant record with known event times, pauses and missing responses. It checks that block predictions are unchanged when future observations are appended, that an unavailable task cannot affect a model's state, and that the separate-task model gives the same pleasure update as C when their pleasure components are matched. These are checks of the experimental intervention on information access. They are necessary even if each model's inference routine passes its own numerical checks.

A recording manifest connects each excerpt to its parent waveform, licence record, transformation log and source family. A model manifest connects each committed prediction to its engine arrays, task likelihoods, global-parameter snapshot and personal-state version. The scoring program reads committed predictions and raw response records without refitting. This separation permits an independent reviewer to reconstruct what was available before each scored block.

5. Confirmatory analysis

Primary estimand

Let $\hat{y}_{utm}$ be model m 's predictive median for listener u at eligible time t , and let y_{ut} be the observed pleasure. The median is used because the primary loss is absolute error. For each participant,

average the time-point errors within each valid scored block and then average the blocks with equal weight:

$$e_{um} = \frac{1}{|\mathcal{B}_u|} \sum_{b \in \mathcal{B}_u} \frac{1}{|V_{ub}|} \sum_{t \in V_{ub}} |y_{ut} - \hat{y}_{utm}|. \quad (12)$$

Here $\mathcal{B}_u$ is the prespecified set of valid blocks among 2–6, and V_{ub} is their common valid-time mask across all models. With complete data, $|\mathcal{B}_u| = 5$. Minimum block and participant coverage are fixed before recruitment; they cannot be chosen after inspecting model errors. The target contrast is

$$\Delta = \frac{1}{N} \sum_{u=1}^N (e_{uC} - e_{uR}). \quad (13)$$

Positive values favour relational prediction. Equal participant weighting prevents long or densely sampled records from dominating the estimate. The participants, rather than audio frames, are the independent units for the primary uncertainty calculation.

A participant-resampled 95% confidence interval accompanies Δ . The procedure resamples each participant's complete pair of errors and preserves their internal time dependence. This inference describes the sampled participants under the selected music panel. A secondary analysis accounts for source-family clustering; generalization to new music requires the held-out-family and replication tests described below.

The smallest effect of practical interest, δ^* , is expressed in pleasure-scale error units and fixed before collection. An interval entirely above zero supports an average predictive gain under the panel; an interval entirely above δ^* supports the stronger claim of practically useful improvement. An interval contained within $[-\delta^*, \delta^*]$ is compatible with practical equivalence under the chosen interval rule. Other outcomes remain inconclusive about these claims. Statistical detection and practical usefulness should both be reported.

Timing, missingness and calibration

The response grid and any correction for reporting delay are chosen using development or pilot data and then held fixed. Optimizing a lag against each evaluated trajectory would let the target determine its own alignment. The primary mask excludes known interruptions and unavailable responses using logged events and prespecified rules. Target pleasure is not imputed to manufacture a complete record. The reasons for missingness and exclusions are reported by participant and block.

Normalization parameters, feature selection, response smoothing and all model tuning are estimated within development data. The same mask and scoring grid apply to every comparator. A sensitivity analysis can examine a small, predefined range of lag assumptions, but it must display all of them rather than replace the primary result with the most favourable alignment.

Absolute error addresses the accuracy of the central forecast. It does not assess whether its uncertainty is trustworthy. Secondary analyses therefore examine predictive interval coverage, interval width and a proper distributional score compatible with the censored observation likelihood. Excessive confidence after auxiliary updates would count against the proposed model even if median error improved slightly. Calibration should also be examined by block, since an accumulating posterior can become overconfident as dependent reports are mistakenly counted as new evidence.

History gain (B versus A), direct-feedback gain (C versus B), separate-task and yoked controls, and performance across source

families are secondary comparisons. A prespecified family of inferential secondary contrasts receives a Holm adjustment. Additional descriptive plots remain labelled as such. A mixed-effects or multiway-resampling sensitivity analysis may address crossed participant and recording effects, but it must not treat time points as independent replications of the study.

6. Discovery beyond pleasure-labelled dimensions

A separate exploratory model

The confirmatory comparison uses a small, theory-motivated set of task relations. A broader analysis asks whether other MI dimensions are associated with pleasure after accounting for known predictors. Large feature libraries and systematic association searches have precedents in time-series phenotyping and phenome-wide analysis (Fulcher et al., 2013; Fulcher and Jones, 2017; Denny et al., 2013). MI could supply an inspectable search space, but the existence of that space would not make the search method itself new.

A schematic candidate model is

$$\mathbb{E}[Y_{ust}] = b_u + f_0(\mathbf{z}_{st}, \mathbf{c}_{ust}) + \sum_j \gamma_j \beta_{uj} \phi_j(\mathbf{z}_{st}, \mathbf{c}_{ust}), \quad (14)$$

where f_0 contains the locked baseline, ϕ_j specifies a candidate transformation or relation, γ_j selects it, and β_{uj} allows its contribution to vary across listeners. The expression is a search specification, not a fitted result. Candidate lags, nonlinearities, interactions and subgroup terms must be counted as part of the search rather than described as free follow-up checks.

The catalogue can include a coordinate whose current label seems remote from pleasure. For example, modulation variability might interact with familiarity in predicting an individual's response. This is a candidate to test, not a finding. It could reflect an acoustic preference, a correlated musical structure, a task artefact or a variable already represented elsewhere. The first question is whether it adds reproducible prediction after the baseline and relevant context have been considered.

Dependence, confounding and level of analysis

Figure 4 illustrates several ways an apparently simple correlation can mislead. A symmetric quadratic relation has zero Pearson correlation despite complete dependence. Conversely, an association pooled across listeners can have the opposite sign within every listener. A population relationship therefore cannot automatically be interpreted as a personal update rule (Fisher et al., 2018). Within-person deviations and between-person differences should be modelled separately where the design supports both.

A shared cause can also create a relation with no unique contribution from the candidate. For independent unit-variance Z , ϵ_X and ϵ_P , set $X = Z + \epsilon_X$ and $P = Z + \epsilon_P$. The marginal correlation is 0.5, whereas the residual covariance after conditioning on Z is zero. In music, intensity, familiarity, order or recording identity may play analogous roles. Adjustment does not prove that all confounding has been removed, but it makes the assumed comparison explicit.

The engine's dimensions are themselves dependent. Some share acoustic inputs, reuse temporal arrays or combine other outputs deterministically. Consequently, 97 R^3 channels and 889 C^3 dimensions are not 986 independent opportunities for discovery. The archive should retain feature lineage, correlations and algebraic relationships. Where several candidates represent nearly the

same signal, results should be reported as a related set rather than as multiple independent discoveries.

Selection, multiplicity and replication

The exploratory search uses development participants and development source families. Every candidate transformation is fitted inside the relevant training split. Selection on a dataset followed by a significance test on the same selected effect would be circular (Kriegeskorte et al., 2009). Nested selection constrains fitting, but it cannot undo earlier design choices informed by a supposedly held-out sample.

Sparse priors can make the cost of searching many relations explicit (Scott and Berger, 2010). If a candidate has prior inclusion probability π and Bayes factor BF, then

$$\Pr(\gamma = 1 \mid D) = \frac{\pi \text{BF}}{1 - \pi + \pi \text{BF}}. \quad (15)$$

With $\pi = 0.01$ and $\text{BF} = 100$, this probability is approximately 0.503. A prior of exactly zero would prohibit discovering the relation at all. Sensitivity to a plausible range of sparse priors should therefore accompany the search. Shrinkage is useful for estimation, but it does not substitute for an independent test of a selected relation.

For frequentist confirmation, the family of selected hypotheses and its error criterion are declared before replication. Benjamini–Hochberg requires its applicable dependence conditions; Benjamini–Yekutieli provides a more conservative option under arbitrary dependence when the individual p -values are valid (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). Neither correction repairs invalid inference caused by treating serially dependent frames as independent observations. The resampling or model-based test must first respect the sampling structure.

Stability across clustered resamples can identify relations worth following up (Meinshausen and Bühlmann, 2010). Before confirmation, freeze the feature definition, lag, direction or shape, context interaction, preprocessing and effect-size criterion. Test that specification in new listeners and new source families. A failed replication should remain visible, alongside the development estimate and the search that produced it. Even a replicated predictor remains an association until an appropriate intervention supports a causal interpretation.

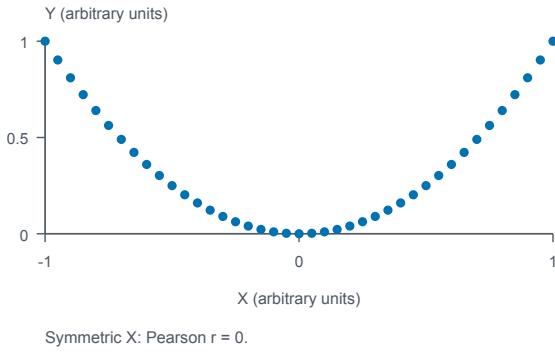
7. Validation and sample-size planning

Separate listeners from source families

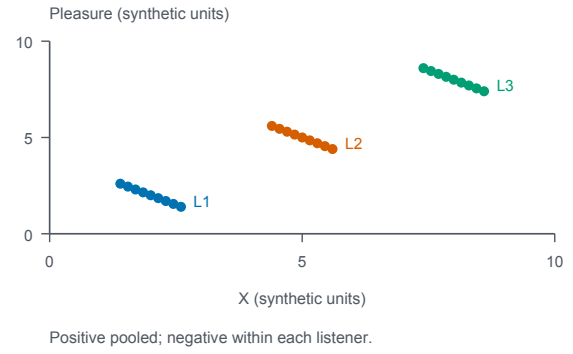
A new participant and a new recording are different generalization problems. The evaluation therefore distinguishes known and new listeners crossed with known and new source families (Figure 5a). Shared model parameters are fitted only in the development partition. A new listener may accumulate a personal state through the same block schedule, but their future blocks cannot influence earlier predictions. New-family evaluation keeps transformed versions and near-duplicate excerpts with their parent family.

The primary R-versus-C contrast tests relational updating under the chosen representation. Additional controls ask whether the result depends on MI's composites: an R^3 -only model, simple history or running-mean calibration, and a generic joint model with comparable capacity. A positive R-versus-C difference without an advantage over these comparators would support the use of auxiliary evidence, while leaving the specific value of the selected MI computations unresolved.

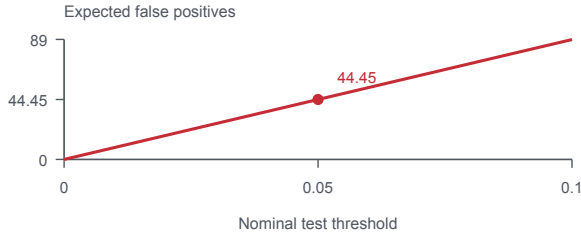
a Nonlinear dependence



b Pooled versus within-listener association



c Uncorrected search across 889 nulls



d Evidence required for a sparse relation

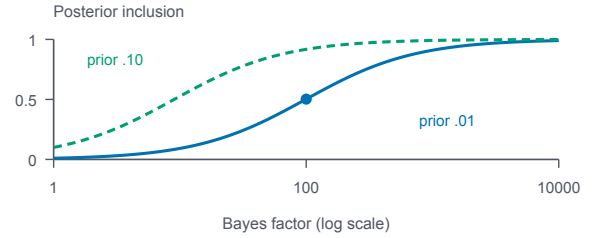
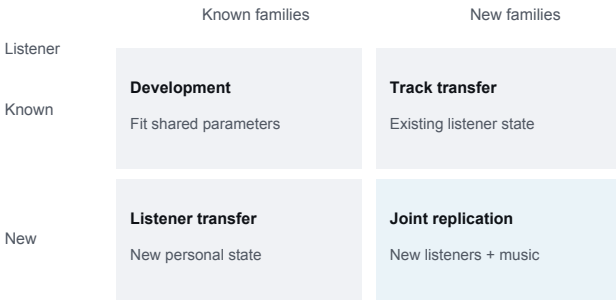


Figure 4. Why an unexpected association needs a separate analysis. (a) Symmetric samples from $Y = X^2$ have zero Pearson correlation despite deterministic dependence. (b) Synthetic observations from three listeners have a positive pooled association and a negative association within each listener. (c) Testing 889 null hypotheses at an uncorrected threshold of 0.05 yields 44.45 expected false positives when each test has that size, regardless of dependence among tests. (d) With prior inclusion probability 0.01, a Bayes factor of 100 gives posterior inclusion probability about 0.503; the dashed curve uses a prior of 0.10. The plots are analytic or synthetic diagnostics and contain no participant measurements.

a Listener and source-family partitions



Transforms and excerpts stay with their source family.

b Illustrative paired-normal sensitivity

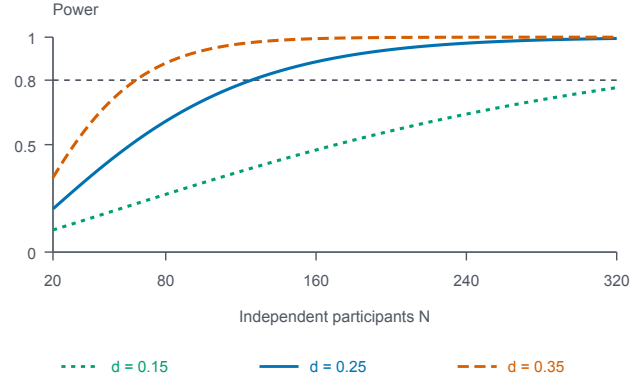


Figure 5. Generalization and planning. (a) Listener and source-family partitions distinguish fitting from transfer. The lower-right cell requires both new listeners and new recording families. (b) Known-standard-deviation paired-normal power for two-sided $\alpha = 0.05$, standardized paired effects $d = 0.15, 0.25, 0.35$ and 20–320 independent participants. The horizontal guide is 0.80. These curves illustrate sensitivity to assumptions; they neither estimate an MI effect nor determine the study's final sample size. Final planning requires simulations of the actual block structure, missingness and crossed participant–recording variation.

Once a test partition has repeatedly influenced model or feature choices, it becomes development data for subsequent claims. Holding out rows in a file is insufficient if their results guide iteration (Dwork et al., 2015). The study record should show when each partition was created, when its outcomes became accessible and which choices followed that access. An external replication provides a clearer boundary than repeatedly recycling one nominal test set.

Planning around an interpretable effect

Sample size should reflect the primary paired contrast and desired precision around a scientifically relevant effect (Lakens, 2022).

Figure 5b uses a paired-normal test with known standard deviation. For standardized paired effect d and N independent participants, power at two-sided $\alpha = 0.05$ is

$$\mathcal{P}(N, d) = 1 - \Phi(z_{0.975} - d\sqrt{N}) + \Phi(-z_{0.975} - d\sqrt{N}). \quad (16)$$

Its assumptions are deliberately simpler than the proposed experiment. The curve should not be used as though the effect or variance had already been observed in MI participants.

A pilot estimates task burden, response coverage, plausible reporting delay and the variability of the participant-level paired difference. These estimates inform a range of generative simulations rather than a single optimistic effect. The simulations vary listener heterogeneity, source-family effects, serial dependence, missingness and the strength or sign of cross-task relations. The final choice of N , attrition allowance, δ^* and stopping rule is frozen before confirmatory recruitment.

The study should also be informative when the relation is absent. Simulations with a zero bridge test false-positive behaviour and the calibration of the interval for Δ . Simulations with a reversed relation test whether the model becomes confidently wrong or appropriately limits transfer. Parameter recovery alone is insufficient: models can recover different parameterizations while making nearly indistinguishable predictions. Both parameter and model recovery, alongside predictive calibration, are needed.

Simulation conditions and diagnostics

The simulation matrix includes at least three relation regimes: no cross-task dependence, dependence with the fitted sign, and dependence opposite to the fitted sign. Each is crossed with low and high auxiliary-response noise, weak and strong between-listener heterogeneity, and realistic ranges of within-excerpt dependence. Additional conditions examine uncertain history and non-random loss of responses. Parameter values should cover pilot-compatible ranges and deliberately adverse cases, with the generative assumptions reported.

For each condition, save the true states, generated observations, committed predictions, posterior summaries and participant-level errors. Compare the nominal and observed interval coverage for Δ , the frequency of a positive claim under the null, the bias and dispersion of the paired effect, and the frequency with which the procedure selects an incorrect relation. Recovery plots must include failed or boundary fits. Discarding those runs would conceal the conditions in which the proposed inference is unreliable.

For an estimated simulation probability $\hat{p}$ based on M independent replications, report the Monte Carlo standard error

$$\text{MCSE}(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{M}}. \quad (17)$$

Choose the number of replications so that simulation error is small relative to the decision being made. A visually smooth power curve is not enough. If independent replications are expensive, a coarse exploration can identify the relevant region, followed by a larger fixed simulation at the proposed operating point.

Items to freeze before confirmatory collection

The preregistration must identify the primary representation, state dimension, task likelihoods, priors, global fitting data and numerical update procedure. It must also specify the stimulus-family partition, participant criteria, session schedule, common validity mask, lag treatment, minimum coverage, primary contrast, interval construction, practical-effect threshold and stopping rule. Sensitivity analyses should be listed with their complete ranges. Choices still unresolved in this draft must not be presented as if they had already been preregistered.

The discovery analysis needs a separate record of feature families, transformations, lags, interactions, selection method and multiplicity handling. After development, a dated confirmation specification names the selected candidates and the independent participants and recordings on which they will be tested. Changes made after seeing confirmation data become explicitly exploratory. This record makes an unexpected association assessable without requiring that every possible relation was anticipated in the initial theory.

8. Adaptive questioning as a subsequent experiment

The initial study fixes task exposure so that all virtual models can be compared on the same record. If relational updating proves

useful, a subsequent experiment can ask whether the next question should depend on the current posterior. Bayesian preference elicitation and goal-oriented experimental design provide relevant precedents (Guo and Sanner, 2010; Attia et al., 2018). For a query q and possible response R_q , one prediction-oriented rule is

$$q^* = \arg \max_q \{I(P^*; R_q | D, q) - \lambda \text{cost}(q)\}. \quad (18)$$

The objective is uncertainty about a specified future pleasure target, not uncertainty about every model parameter. Cost can include time and burden, with its definition and trade-off fixed in advance. A question that is highly informative about an unrelated parameter need not be useful for this target.

A discovery objective would instead concern a candidate relation set G . It could prioritize observations that distinguish a linear relation from a context-dependent or absent one. Here G denotes statistical alternatives; calling it a causal graph would require additional assumptions and interventions. Prediction and discovery objectives may select different questions and should be evaluated separately.

The subsequent experiment would randomize participants, at equal total query budgets, to direct-pleasure querying, a static mixture of tasks or an adaptive mixture. Evaluation music and the primary pleasure-scoring procedure remain fixed outside the policy's choices. This avoids defining success only on items the policy preferred to ask about. Feedback-driven changes in musical discovery provide a related application context (Nonnemaker et al., 2025), but recommendation success and the mechanism of pleasure require their own outcome definitions.

Asking can change listening. A continuation probe may direct attention, increase familiarity or alter the interpretation of later music. Replaying alternative update rules over a completed record cannot estimate the causal effect of having asked a different question. Randomized policy assignment is therefore needed for that claim. A further intervention could manipulate a frozen candidate acoustic feature while matching relevant alternatives and measuring off-target changes; reproducible prediction would guide that experiment without completing its causal argument (Peters et al., 2016).

9. Discussion

The proposed experiment tests a limited but consequential idea: information about pleasure may be obtained through observations that do not directly ask about pleasure. A familiarity report can constrain one interpretation of a response; an expectancy choice can constrain another. When those constraints are informative under a well-specified joint model, they change the prediction of future pleasure. The mathematical example makes that possibility explicit and also shows why a mistaken relation can be harmful.

MI's potential contribution lies in making the representation and assumed relations available for inspection. An investigator can identify the acoustic inputs to a candidate, remove a connection, replace a composite with its constituents, or ask which observed response changed a personal prediction. This organization may make failures easier to diagnose. The experiment must establish whether it also provides predictive or scientific value beyond simpler alternatives.

The broader discovery aim is compatible with this approach, provided the search remains open to unexpected dimensions and equally open to null results. A feature without an existing pleasure label could become a useful candidate if it adds prediction across listeners and music. Its present label neither disqualifies

it nor supplies a mechanism. The relevant contribution would be the reproducible relation, its conditions and the experiment it motivates.

Several limits remain. The audio representation contains interdependent, historically developed hypotheses. History can be incomplete or biased, continuous reporting can alter experience, and different observation models can explain the same response pattern. A compact personal state may be inadequate for context-sensitive preferences. Conversely, a very flexible state may fit noise or obscure the distinction between tasks. These risks motivate staged recovery, matched comparators and a separate replication rather than a broad claim of personalization from one demonstration.

The protocol also preserves a distinction between updating the model and changing the person. A model may learn to describe a stable preference without the listener learning anything new. To study listener learning, an experiment would manipulate exposure or transition histories and measure expectation and pleasure separately for matched probes. Computational memory models such as PPM-Decay offer one relevant way to formulate temporal hypotheses (Harrison et al., 2020); their predictions would still require an observation design that separates exposure, expectation and reward.

At present, the personal posterior, its task interfaces and the confirmatory experiment are proposed work. Source inspection supports the architectural counts and selected routes shown here. Analytic calculations support the conditional statements about information transfer. Neither is evidence that participants benefit from

the proposed model. That evidence would come from the committed predictions, their calibrated uncertainty and replication under the boundaries specified in this protocol.

Data and code availability

No participant dataset accompanies this protocol. Editable figure-generation code, analytic inputs and the manuscript source accompany this edition. The source audit records hashes for the inspected engine files and the selected subgraph; it is an inspection record, not an end-to-end experimental validation. The existing MI architecture and its research-access arrangements are described in the companion architecture manuscript. Future releases should include de-identified response records where consent permits, stimulus manifests, partition identities, committed predictions and scoring code. Licensed waveforms and private listening histories require their own access arrangements.

Author statements

The author designed the proposed study and is responsible for its scientific content. No external funding supported this manuscript. Human recruitment and data collection require prior ethics approval and informed consent, including arrangements for withdrawal and protection of listening-history data. No such approval or completed collection is claimed here. Claude Code and Codex assisted code development, drafting, document preparation and computational checks; these tools are not authors and do not replace author review.

References

- Attia A, Alexanderian A, Saibaba AK (2018). Goal-oriented optimal design of experiments for large-scale Bayesian linear inverse problems. *Inverse Problems* 34, 095009. [Source](#)
- Benjamini Y, Hochberg Y (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc B* 57, 289-300. [Source](#)
- Benjamini Y, Yekutieli D (2001). The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29, 1165-1188. [Source](#)
- Bonilla EV, Chai KMA, Williams CKI (2007). Multi-task Gaussian Process Prediction. *NeurIPS* 20. [Source](#)
- Cheung VKM et al. (2024; online 2023). Cognitive and sensory expectations independently shape musical expectancy and pleasure. *Philos Trans R Soc B* 379, 20220420. [Source](#)
- Denny JC et al. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 31, 1102-1111. [Source](#)
- Dwork C et al. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science* 349, 636-638. [Source](#)
- Fisher AJ, Medaglia JD, Jeronimus BF (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *PNAS* 115, E6106-E6115. [Source](#)
- Fulcher BD, Jones NS (2017). hctsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction. *Cell Syst* 5, 527-531.e3. [Source](#)
- Fulcher BD, Little MA, Jones NS (2013). Highly comparative time-series analysis: the empirical structure of time series and their methods. *J R Soc Interface* 10, 20130048. [Source](#)
- Guo S, Sanner S (2010). Real-time Multiattribute Bayesian Preference Elicitation with Pairwise Comparison Queries. *AISTATS, PMLR* 9, 289-296. [Source](#)
- Harrison PMC et al. (2020). PPM-Decay: A computational model of auditory prediction with memory decay. *PLoS Comput Biol* 16, e1008304. [Source](#)
- Janata P (2009). The neural architecture of music-evoked autobiographical memories. *Cereb Cortex* 19, 2579-2594. [Source](#)
- Kriegeskorte N et al. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci* 12, 535-540. [Source](#)
- Lakens D (2022). Sample Size Justification. *Collabra: Psychology* 8, 33267. [Source](#)
- Mas-Herrero E et al. (2014). Dissociation between musical and monetary reward responses in specific musical anhedonia. *Curr Biol* 24, 699-704. [Source](#)
- Meinshausen N, Bühlmann P (2010). Stability selection. *J R Stat Soc B* 72, 417-473. [Source](#)
- Nonnemaker A et al. (2025). Feedback-Driven Gradual Discovery for Expanding Musical Preferences. *RecSys*, 605-609. [Source](#)
- Peters J, Bühlmann P, Meinshausen N (2016). Causal inference by using invariant prediction: identification and confidence intervals. *J R Stat Soc B* 78, 947-1012. [Source](#)
- Rasmussen CE, Williams CKI (2006). *Gaussian Processes for Machine Learning*. MIT Press. Chapter 2. [Source](#)
- Scott JG, Berger JO (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann Stat* 38, 2587-2619. [Source](#)
- Turner BM et al. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage* 72, 193-206. [Source](#)
- van den Bosch I, Salimpoor VN, Zatorre RJ (2013). Familiarity mediates the relationship between emotional arousal and pleasure during music listening. *Front Hum Neurosci* 7, 534. [Source](#)
- Wilson RC, Collins AGE (2019). Ten simple rules for the computational modeling of behavioral data. *eLife* 8, e49547. [Source](#)