

Learning the Listener

A relational Bayesian experiment for musical pleasure, personal memory and discovery across dimensions

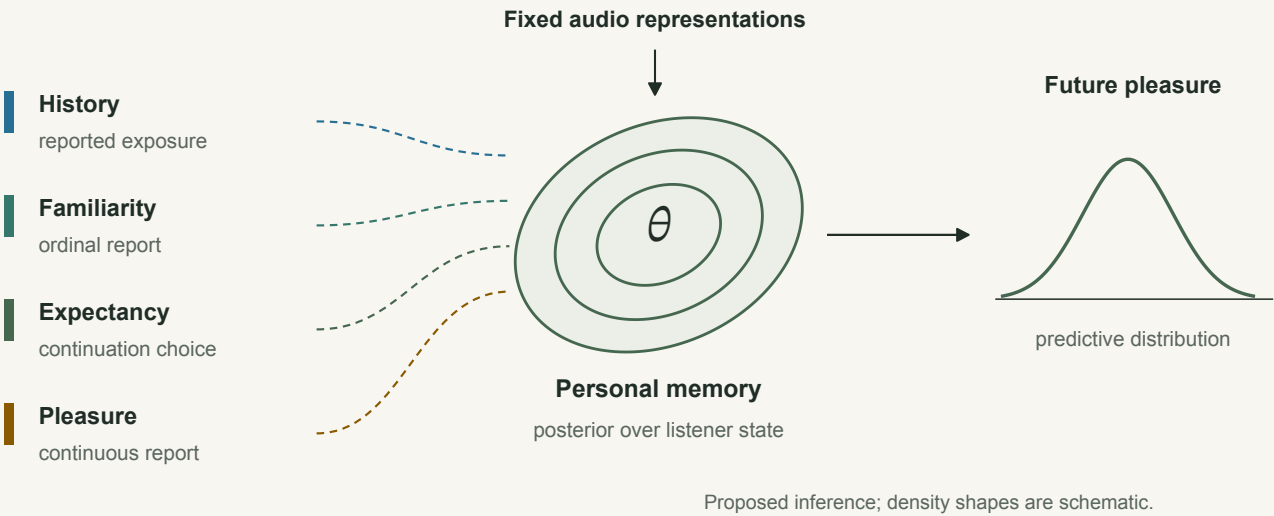
Amaç Erdem Independent researcher

Abstract

Can responses outside pleasure improve a personal model of musical pleasure? We propose an experiment that holds the Musical Intelligence (MI) audio engine fixed while updating an explicit Bayesian listener state. Identical listening episodes support direct-feedback and relational models; familiarity and expectancy observations test whether shared structure adds predictive information.

A separate discovery stream searches eligible MI dimensions for unanticipated associations, then freezes candidates for independent confirmation. The contribution is a traceable architecture for prediction, measurement and hypothesis generation. No human results are reported.

a The experimental argument



Which observation changes which prediction - and why?

- | | | |
|---|--|---|
| 01 Personalize
Test transfer beyond direct feedback | 02 Discover
Search beyond pleasure-labelled features | 03 Confirm
Replicate before a new bridge is trusted |
|---|--|---|

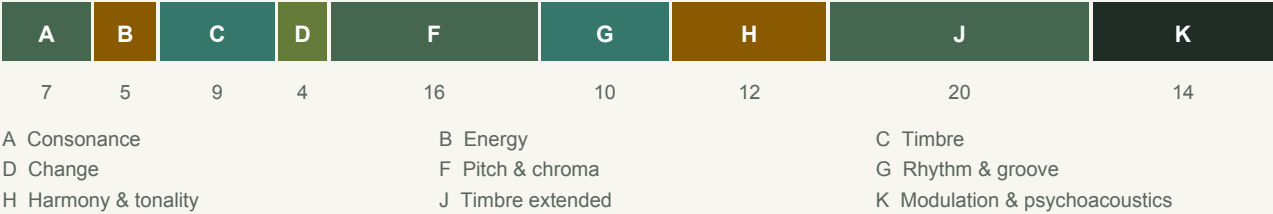
Fig. 1 Three linked aims, three distinct evidential standards. Dashed links are proposed statistical relations, not measured neural connections. The design extends a source-audited audio architecture; it does not demonstrate successful personalization, biological validity or a discovered mechanism. [1-6]

A fixed engine, a changing listener

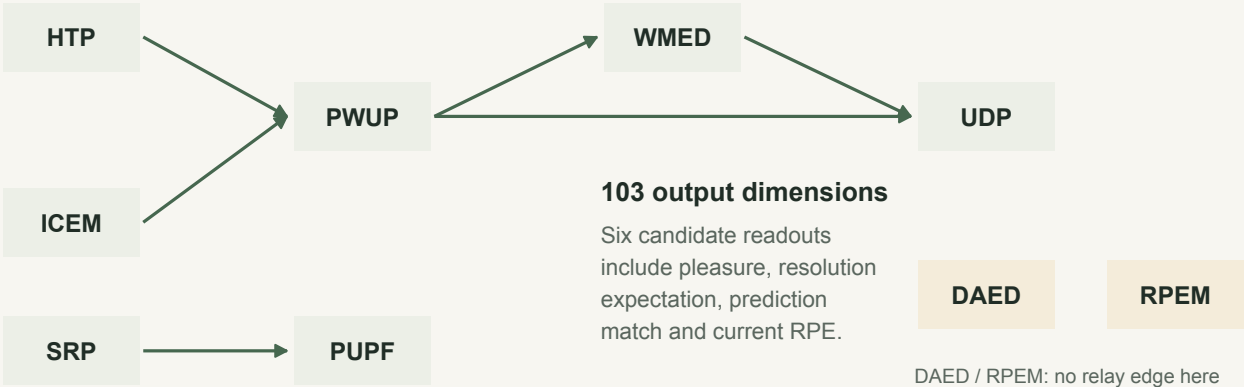
Solid structure is source-audited. The dashed personal layer is a proposed downstream extension.

a An interpretable audio coordinate system

97 R3 channels / 9 groups; prepared temporal representations supply T3 inputs.



b Audited pleasure subgraph: 9 mechanisms, 6 declared relay edges



c Proposed personal layer

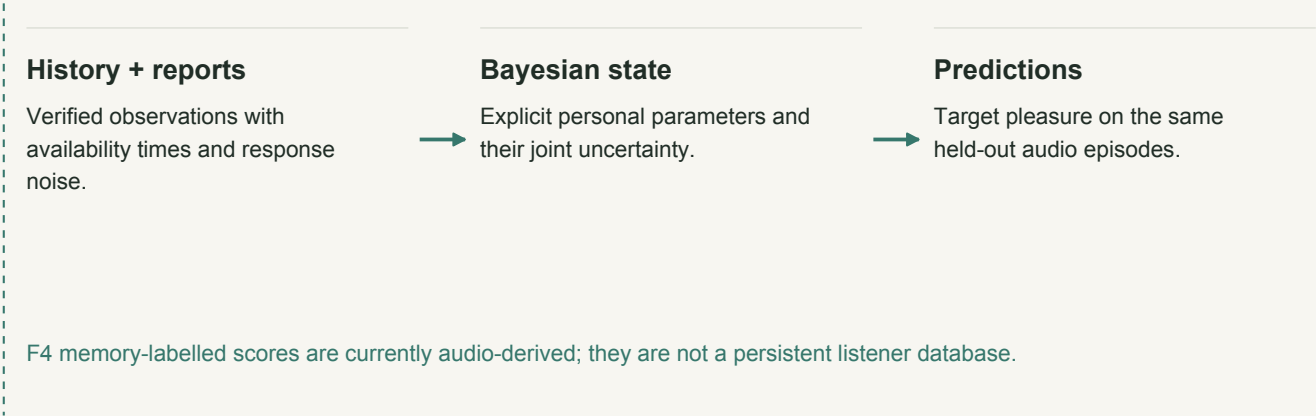
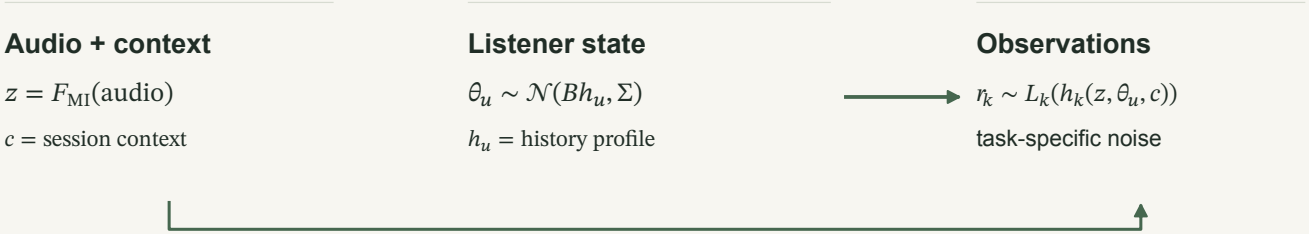


Fig. 2 Source audit: the existing figure contract records this selected subgraph. The broader registry is a discovery dictionary, not a set of independent measurements. Inspection of HCMC.compute shows audio/temporal/relay inputs and 11 bounded outputs, with no participant-response or persistent personal-state interface. Counts and hashes are pinned in the source ledger. [S1-S3]

A response is an observation

The model must distinguish the music, the listener, the measurement and the target.

a Proposed hierarchical generative model



$$p(\theta_u \mid D, r_k) \propto p(r_k \mid \theta_u, z, c) p(\theta_u \mid D)$$

State uncertainty is integrated; bridge uncertainty is assessed with frozen-draw sensitivity analyses.

b Two primary auxiliary tasks; separate likelihoods

Pleasure Primary target	0-100 continuous rating Censored Gaussian around $100 \times \text{sigmoid}(\eta_P)$
Familiarity Auxiliary task 1	1-5 ordinal report Ordered-logit thresholds; no equal-step assumption
Expectancy Auxiliary task 2	Choice between two continuations Bernoulli / two-option softmax

Recognition accuracy, surprise and urge-to-move are optional extensions, not interchangeable outcomes.

c A compact state before a broad search

A candidate θ_u contains familiarity and expectancy sensitivities plus hedonic intercept and weights. Development data constrain a small shared covariance; individual data update the posterior. The exact dimension, priors, link functions and residual dependence are frozen after recovery checks.

Primary: infer a stable session-level state. Time-varying latent preferences are a separate sensitivity model.

Fig. 3 The familiarity task measures subjective prior familiarity; a continuation choice measures relative expectation. Neither is treated as a pleasure label. A low-pleasure response can coexist with intact music perception, so a universal positive bridge is not imposed. Global parameters are fitted only in development participants. The personal update estimates state; it does not by itself prove that the listener learned. [1-5,7]

When another dimension helps

An exact Gaussian example proves a conditional possibility, not an empirical guarantee.

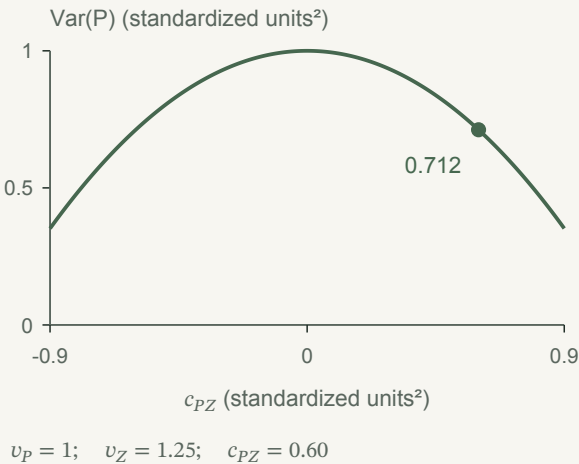
a Condition on all information already in D

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

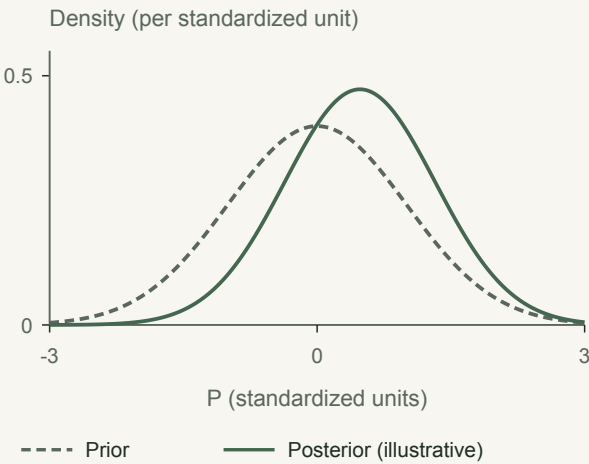
$$\text{Var}(P \mid D, Z) = v_P - \frac{c_{PZ}^2}{v_Z}$$

P: future pleasure; Z: auxiliary response including measurement noise. Joint Gaussianity and known conditional covariance are assumed. Negative covariance can be just as informative as positive covariance.

b Conditional uncertainty



c A target posterior after Z = 1



d Precision is only useful if the bridge is right

No auxiliary update		1.000
Correct bridge		0.712
Wrong sign of bridge		1.864

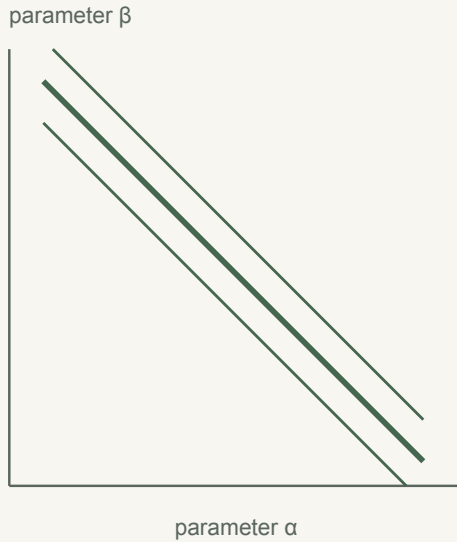
Expected squared prediction error under the same true model; analytic values, not participant results.

Fig. 4 For several auxiliary observations, the variance reduction is $c^T V^{-1} c$ and their full covariance matters. A duplicated observation adds no new information. With uncertain bridges or non-Gaussian models, a realized observation need not reduce variance. Expected information is nonnegative under a coherent model, but accuracy under model misspecification can deteriorate. The wrong-sign example exposes that distinction. [1,2,8]

More answers, or better questions?

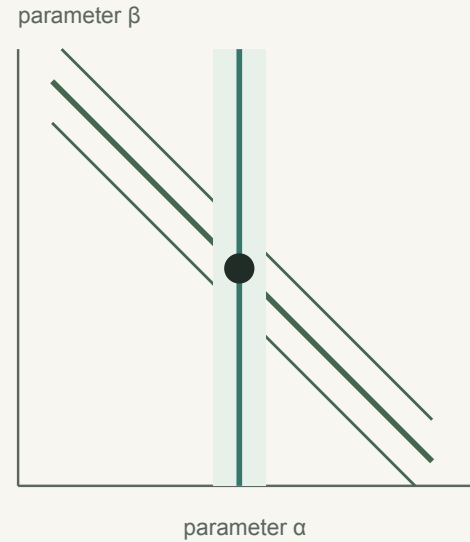
A useful task separates explanations that pleasure reports alone cannot distinguish.

a Pleasure alone



$P = \alpha + \beta$; design rank = 1
Different parameter pairs explain the same P .

b Pleasure + auxiliary



$P = \alpha + \beta$; design rank = 2
 $Z = \alpha$ adds a non-redundant projection.

c The information test

$$I(P^\star; Z \mid D) = \mathbb{E}_{Z \mid D} [\text{KL}(p(P^\star \mid D, Z) \parallel p(P^\star \mid D))] \geq 0$$

This is expected information under the specified model. For squared loss, the expected reduction in Bayes risk equals $\text{Var}(\mathbb{E}[P^\star \mid D, Z] \mid D)$. It does not promise lower error for every listener or a particular observed response.

d A transformation is not an additional observation

$$Q = g(X) \implies I(P^\star; Q \mid X, D) = 0$$

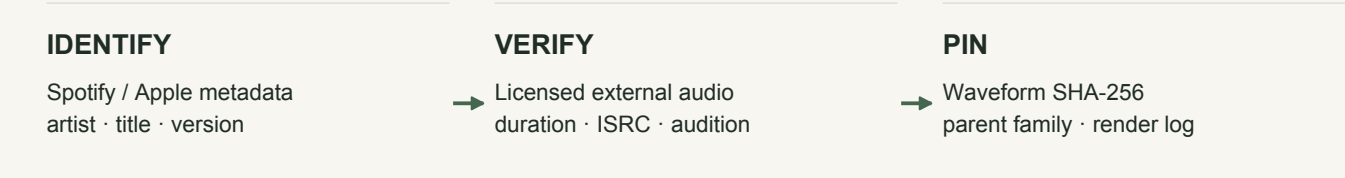
A composite MI score can still supply a valuable representation or inductive bias. Its deterministic derivation must remain visible; re-counting it as independent evidence would exaggerate certainty.

Fig. 5 The upper panels depict likelihood constraints in an idealized two-parameter model, not fitted human posteriors. A prior can regularize an unidentifiable model without adding data identification. Recovery simulations must show that the selected tasks can distinguish the intended latent quantities before they are given psychological interpretations. [7,8]

One recording, one provenance chain

Titles identify candidates. The actual waveform and exposure evidence define the observation.

a From catalogue identity to an analyzable stimulus



A matching name is insufficient. Live takes, remixes and edited excerpts retain distinct waveform identities.

b Exposure is evidence with a source and a cutoff

- OBSERVED Voluntary history export or self-report with timestamps, source and missingness.
- INFERRED Probabilistic exposure and familiarity profile; uncertainty remains explicit.
- NOT ESTABLISHED Playlist membership alone does not prove listening, liking or autobiographical recall.

c A crossed stimulus ledger

SOURCE FAMILY	Original excerpt	Transformed version	Repeated excerpt
Family 01	01.1.1	01.1.2	01.1.3
Family 02	01.2.1	01.2.2	01.2.3
Family 03	01.3.1	01.3.2	01.3.3

All derivatives stay inside their source-family partition. Transformed repeats are not new-track transfer.

Fig. 6 The provisional design uses participant-selected familiar music and unfamiliar families sampled independently of the evaluation models. A fixed rendered-audio manifest stores boundaries, operators, units, playback level, engine version and array hashes. Novelty is checked by participant report. History, familiarity and pleasure retain different evidential roles; private narrative content is unnecessary. [3,6,9,S1]

The same listener, the same music

Virtual model conditions share every scored episode; only their permitted information differs.

a Six provisional blocks; updates occur only between blocks



Each block: commit predictions → play common audio → collect reports → close → update.

b Frozen core, controlled access to evidence

MODEL	History	Pleasure	Auxiliary	Bridge	Role
A Population	-	-	-	-	baseline
B History	yes	-	-	-	secondary
C Direct feedback	yes	past	-	off	primary
C⊥ Separate tasks	yes	past	past	off	diagnostic
R Relational	yes	past	past	on	primary
R~ Yoked responses	yes	past	yoked	on	control

c A sparse, counterbalanced response schedule

Continuous pleasure accompanies 60 one-minute scored excerpts. Two familiarity prompts and two continuation probes follow selected excerpts per block; probe audio is unscored. Every virtual model faces the same query-induced experience. No within-block report enters a committed prediction.

Provisional burden: 60 min scored audio + about 8 min probes + setup, practice and breaks.

Block 1: B = C = C⊥ = R in their pleasure predictions; a shared history-conditioned marginal prior is required.

Fig. 7 Primary contrast: R versus C, with equal direct-pleasure feedback and identical human burden. C⊥ fits auxiliary heads independently and cannot transfer their evidence to pleasure. Capacity, priors and fitting budget are controlled as far as the models permit; residual differences are disclosed. B-A and C-B retain their original meanings as secondary history and feedback contrasts. Timing and task details require pilot lock. [S1]

A gain must survive the scoring rule

Forecasts are immutable. Participants, not densely sampled slider frames, are the replication unit.

a One confirmatory estimand

$$e_{u,m} = \text{mean}_{b=2,\dots,6} \left[\text{mean}_{t \in V_{ub}} |y_{u,t} - \hat{y}_{u,t,m}| \right]$$

$$\Delta = \text{mean}_u (e_{u,C} - e_{u,R})$$

Positive Δ favors relational prediction on the common 0-100 pleasure scale.

Use the posterior predictive median for MAE. Pair every scored timestamp across models; weight valid blocks equally within listener and listeners equally in the group estimate.

b Commit before observing

1 TRAIN

Only reports available through block b-1.

2 COMMIT

Model hash, array hash, timestamp, cutoff.

3 SCORE

Fixed lag, grid, mask and output mapping.

4 REPORT

Paired Δ and participant-resampled 95% CI.

c Missingness stays visible

Log raw input and audio clocks, pauses, hidden tabs and interruptions. Untouched slider defaults are not measurements.

Use a common valid-time mask. Minimum coverage and participant exclusions are fixed before outcomes are examined. Report loss of data by block and task; do not silently impute target pleasure.

d Separate evidence of improvement from practical usefulness

A 95% interval above zero supports a positive average predictive gain under this design. An interval wholly above a prespecified meaningful difference δ^* supports practical benefit; an interval inside $[-\delta^*, \delta^*]$ supports practical equivalence. Other outcomes remain inconclusive at the chosen precision.

Secondary: calibration / log score / gain by block / history contrasts (Holm) / family-stratified transfer.

The primary interval is conditional on this stimulus panel; independent source-family replication follows.

Fig. 8 The primary population is eligible adult listeners able to complete the tasks under the approved protocol. Language proficiency, hearing criteria, musical background and recruitment quotas are specified before collection. Response lag is estimated in development data only. Mixed-effects and multiway participant/source-family resampling are secondary robustness analyses, not a substitute for an untouched replication panel. [7,10]

Discover beyond the pleasure labels

A broad representation can generate a narrow, falsifiable claim that was not specified in advance.

a **Searchable dictionary, not 986 independent tests**

97

R3 audio channels

T₃

eligible temporal arrays

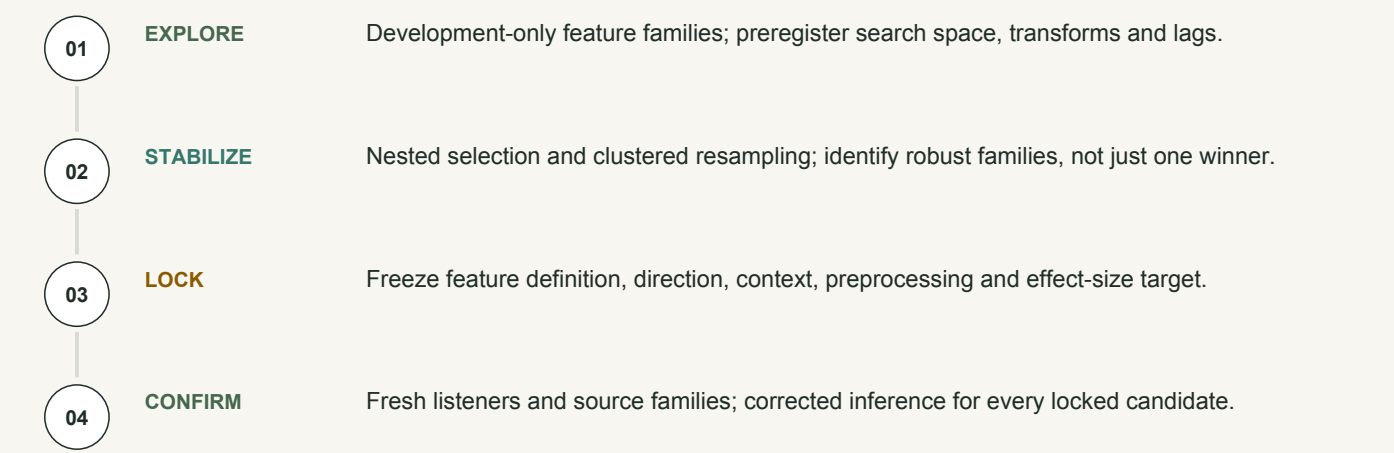
889

C3 dimensions / 84 mechanisms

$$\mathbb{E}[Y_{ust}] = b_u + f_0(z_{st}, c_{ust}) + \sum_j \gamma_j \beta_{uj} \varphi_j(z_{st}, c_{ust})$$

Proposed sparse readout: γ_j selects a relation; β_{uj} varies by listener. Audit dependent features first.

b **A candidate survives four different decisions**



c **The result is initially an association claim**

Example candidate: X = modulation variability → conditional pleasure association

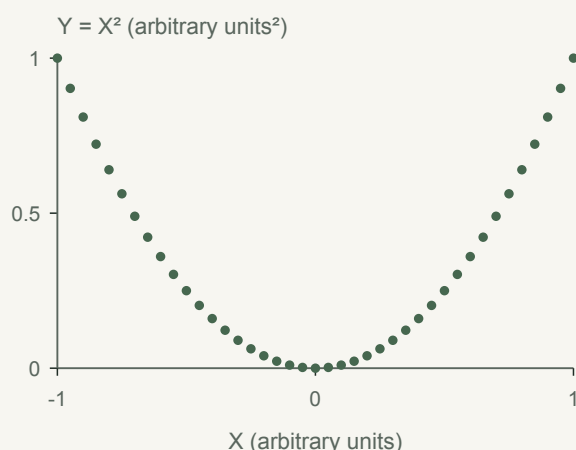
This is an illustrative candidate, not a finding. Ask whether X predicts beyond prespecified acoustics, familiarity and context; test within-person and between-person components separately. Replication may justify a new statistical bridge, not a causal interpretation.

Fig. 9 Feature-based scientific discovery has precedents in comparative time-series analysis and phenome-wide association studies. MI-specific value would lie in the provenance of each dimension, disciplined behavioral sampling and independent confirmation of useful relations. Discovery success is reported separately from the primary predictive test. No architecture label guarantees that a score corresponds to its named psychological mechanism. [11-14,S1]

An unexpected edge needs scrutiny

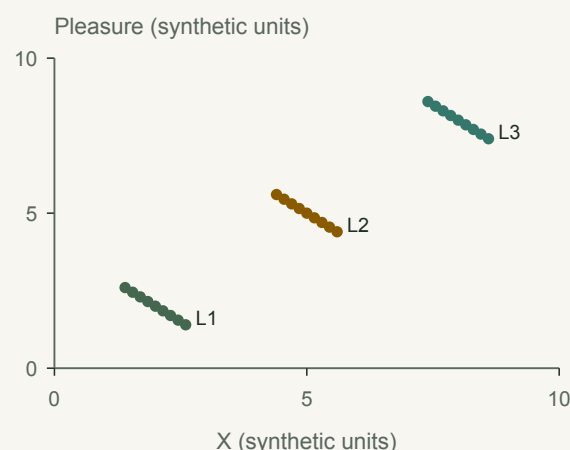
Four exact or synthetic examples show why finding a correlation is only the beginning.

a Zero correlation can miss dependence



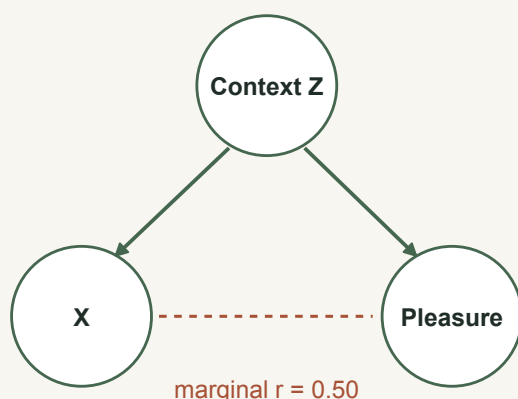
Symmetric X: Pearson $r = 0$; dependence remains.

b Pooled and personal relations can reverse



Three listeners: positive pooled / negative within.

c A third variable can create the edge



$$X = Z + \varepsilon_X; \quad P = Z + \varepsilon_P$$

Independent unit-variance terms give a marginal association and zero residual association after conditioning on Z.

d Multiplicity belongs inside the model

$$889 \times 0.05 = 44.45$$

Expected false positives if 889 true nulls are tested at an uncorrected 0.05 level. Independence is not needed for this expectation.

Posterior edge probability

$$\frac{\pi \cdot \text{BF}}{1 - \pi + \pi \cdot \text{BF}}$$

A prior $\pi = 0$ forbids discovery. A small positive prior permits it; a hierarchical inclusion prior can express sparsity and multiplicity.

$$\pi = 0.01; \quad \text{BF} = 100 \Rightarrow \text{posterior} \approx 0.503$$

Fig. 10 These are constructed examples, not MI or human measurements. Search multiplicity includes transforms, lags, interactions and subgroups. BH requires appropriate dependence conditions; BY tolerates arbitrary dependence given valid p-values. Neither repairs invalid time-series inference. Shrinkage, stability and independent replication address different problems; none should be treated as an automatic certificate of discovery. [15-19]

Transfer must cross a real boundary

Development, locked prediction and replication have different people and different source families.

a Generalization is a two-axis claim

	Known source families	Unseen source families
Known listeners	Development Fit global priors and model choices.	Track transfer New families, existing listener state.
New listeners	Listener transfer New personal state; known library.	Joint replication New listeners AND new families.

No excerpt, transform or near-duplicate source family straddles the declared held-out boundary.

b Controls that can defeat the interpretation

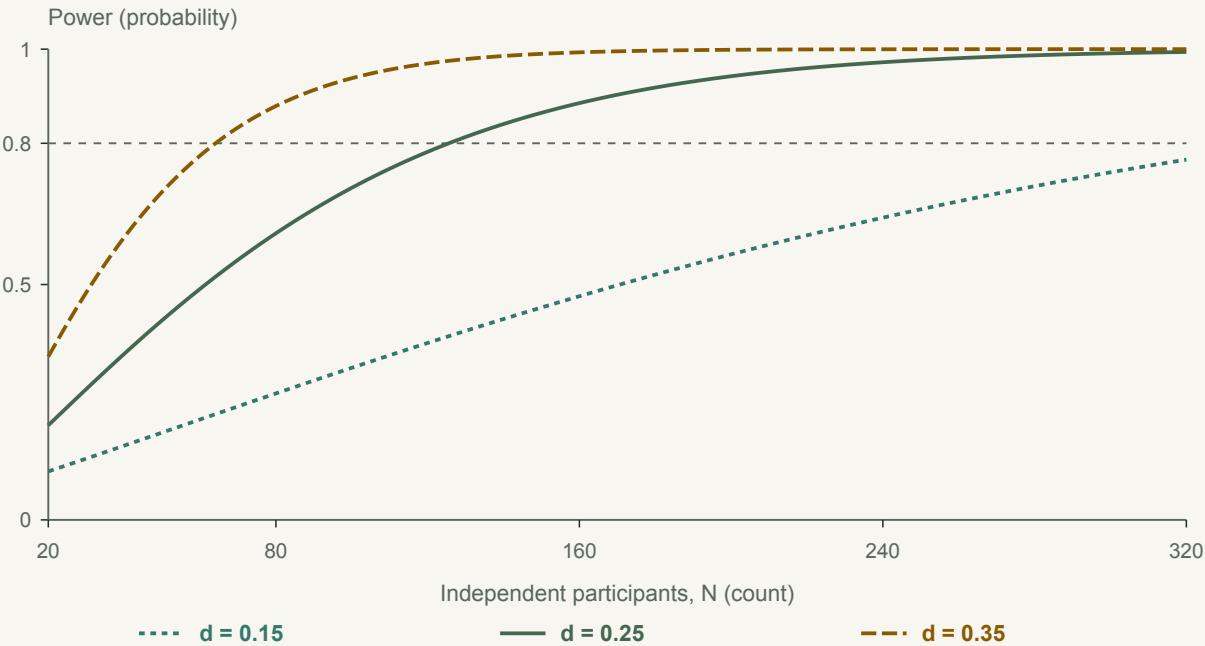
● Bridge disabled	Does the gain disappear when auxiliary-to-pleasure transfer is blocked?
● Yoked auxiliary reports	Do matched but wrong-person responses perform just as well?
● R3-only personal model	Does selected C3 structure add anything beyond low-level audio?
● History-only / running mean	Is the effect merely a stable preference level or response inertia?
● Capacity-matched joint model	Does interpretable structure help beyond generic joint prediction?
● Source-family replication	Does the locked association survive fresh music and fresh listeners?

Fig. 11 All model selection, normalization, feature screening and hyperparameter tuning occur inside development partitions. A later test set becomes training data once it guides choices. Yoking preserves task and block structure; its validity and exchangeability are checked. An online personal update may use preceding blocks from a new listener, but population parameters and scored forecasts remain locked. [7,14,19,20]

Precision before recruitment

Sensitivity calculations expose assumptions; a full pilot-based simulation sets the final target.

a Illustrative power under a simple paired-normal model



d = mean paired MAE gain / between-participant SD of paired gains. Known-SD normal approximation; independent participants; $\alpha = 0.05$. No human data inform these curves.

b The recruitment target comes from the complete design

DEVELOPMENT PILOT	Estimate variability, response lag, missingness, burden and plausible bridge strengths.
SIMULATE THE PIPELINE	Generate crossed listener/music effects, serial noise, null and beneficial bridges; fit every arm.
RECOVER + CALIBRATE	Check parameter/model recovery, coverage, false-positive rate and wrong-bridge behavior.
LOCK BEFORE RECRUITMENT	Set δ^* , sample size, attrition reserve, exclusions, stopping rule and replication allocation.

Fig. 12 The curves are a transparent sensitivity illustration, not a power claim for MI or a recommended sample size. A fixed confirmatory N is justified from an independent pilot and a prespecified smallest meaningful gain, with sensitivity to variance and dropout. Participant and source-family replication budgets are planned separately. Monte Carlo standard errors are reported for simulation-based design decisions. [7,10]

Ask the question that resolves doubt

Adaptive measurement is a later randomized experiment, after the basic relational model is tested.

a Two legitimate objectives, two different experiments

PREDICTION

$$q^\star = \arg \max_q \{I(P^\star; R_q \mid D, q) - \lambda \text{ cost}(q)\}$$

DISCOVERY

$$q^\star = \arg \max_q \{I(G; R_q \mid D, q) - \lambda \text{ cost}(q)\}$$

$P^\star$ is a prespecified future pleasure target. G indexes candidate statistical relations, not an established causal graph. Acquire information about the declared target rather than shrinking every parameter indiscriminately.

b Randomize people to equal-burden query policies

Direct reports

Fixed pleasure queries per assigned budget.

Static mixed tasks

Fixed pleasure + auxiliary schedule, same time budget.

Adaptive mixed tasks

Model selects among eligible queries, same time budget.

c The claim ladder

Common evaluation music and pleasure scoring remain outside adaptive query choice.

Queries can change attention, familiarity and later pleasure. Randomization estimates the total policy effect; passive replay analyses alone cannot establish the effect of asking.

			Mechanism
Association	Replication	Prediction	A discriminating intervention tests why.
A relation in development data.	A frozen relation survives new data.	It improves locked target forecasts.	

Fig. 13 For causal follow-up, manipulate a frozen candidate dimension while matching prespecified acoustics and testing side effects; a targeted feature edit may change several dimensions. Consent, withdrawal, burden limits and institutional ethics review precede collection. The scientific contribution is a falsifiable measurement architecture: successful transfer, failed bridges and replicated unexpected relations are all informative outcomes. [21-24]

References and provenance

Primary literature informs the design. Local source contracts establish implementation status.

- 1 Turner BM et al. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage* 72, 193-206.
10.1016/j.neuroimage.2013.01.048
- 2 Bonilla EV, Chai KMA, Williams CKI (2007). Multi-task Gaussian Process Prediction. *NeurIPS* 20.
Primary source: NeurIPS
- 3 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.
10.3389/fnhum.2013.00534
- 4 Cheung VKM et al. (2024; online 2023). Cognitive and sensory expectations independently shape musical expectancy and pleasure. *Philos Trans R Soc B* 379, 20220420.
10.1098/rstb.2022.0420
- 5 Mas-Herrero E et al. (2014). Dissociation between musical and monetary reward responses in specific musical anhedonia. *Curr Biol* 24, 699-704.
10.1016/j.cub.2014.01.068
- 6 Harrison PMC et al. (2020). PPM-Decay: A computational model of auditory prediction with memory decay. *PLoS Comput Biol* 16, e1008304.
10.1371/journal.pcbi.1008304
- 7 Wilson RC, Collins AGE (2019). Ten simple rules for the computational modeling of behavioral data. *eLife* 8, e49547.
10.7554/eLife.49547
- 8 Rasmussen CE, Williams CKI (2006). *Gaussian Processes for Machine Learning*. MIT Press. Chapter 2.
Primary source: GPML, chapter 2
- 9 Janata P (2009). The neural architecture of music-evoked autobiographical memories. *Cereb Cortex* 19, 2579-2594.
10.1093/cercor/bhp008
- 10 Lakens D (2022). Sample Size Justification. *Collabra: Psychology* 8, 33267.
10.1525/collabra.33267
- 11 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.
10.1098/rsif.2013.0048
- 12 Fulcher BD, Jones NS (2017). hctsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction. *Cell Syst* 5, 527-531.e3.
10.1016/j.cels.2017.10.001
- 13 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.
10.1038/nbt.2749
- 14 Kriegeskorte N et al. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci* 12, 535-540.
10.1038/nn.2303
- 15 Scott JG, Berger JO (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann Stat* 38, 2587-2619.
10.1214/10-AOS792
- 16 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.
10.1111/j.2517-6161.1995.tb02031.x
- 17 Benjamini Y, Yekutieli D (2001). The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29, 1165-1188.
10.1214/aos/1013699998
- 18 Meinshausen N, Bühlmann P (2010). Stability selection. *J R Stat Soc B* 72, 417-473.
10.1111/j.1467-9868.2010.00740.x
- 19 Fisher AJ, Medaglia JD, Jeronimus BF (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *PNAS* 115, E6106-E6115.
10.1073/pnas.1711978115
- 20 Dwork C et al. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science* 349, 636-638.
10.1126/science.aaa9375
- 21 Guo S, Sanner S (2010). Real-time Multiattribute Bayesian Preference Elicitation with Pairwise Comparison Queries. *AISTATS, PMLR* 9, 289-296.
Primary source: PMLR
- 22 Attia A, Alexanderian A, Saibaba AK (2018). Goal-oriented optimal design of experiments for large-scale Bayesian linear inverse problems. *Inverse Problems* 34, 095009.
10.1088/1361-6420/aad210
- 23 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.
10.1111/rssb.12167
- 24 Nonnemaker A et al. (2025). Feedback-Driven Gradual Discovery for Expanding Musical Preferences. *RecSys*, 605-609.
10.1145/3705328.3748025

LOCAL SOURCE RECORDS

S1 UCSD experiment engineering contract, revision 2026-10-08-v18. S2 MI-test figure contract: selected mechanism graph. S3 R3 group boundaries and HCMC compute interface. Relative paths and SHA-256 hashes accompany this manuscript.

Declarations: no external funding; no participant data collected for this draft. AI assistance: Codex supported drafting, figure code and checks. Author review, ethics approval, implementation, pilot decisions and preregistration remain prerequisites for the proposed study.