

Inspectable representations and auditable inference in computational music neuroscience

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

Abstract. MI-Lab links inspectable music representations, independent observations and auditable numerical analysis. Musical Intelligence (MI) supplies versioned descriptors; personal readouts learn from eligible earlier responses; an LLM assists analysis specification and code. A transitive dependency contract defines admissible information. A proposed benchmark holds MI arrays fixed while varying definitions and provenance, distinguishing documentation effects from representation and analyst effects. Success requires correct or appropriately qualified conclusions, valid information access and rerunnable numerical artifacts. Controls address persuasive labels, leakage, unequal resources and task dependence. No participant, neural or benchmark results are reported.

Prospective framework and evaluation specification. All figures are original schematics; no deployed hybrid service or empirical gain is claimed.

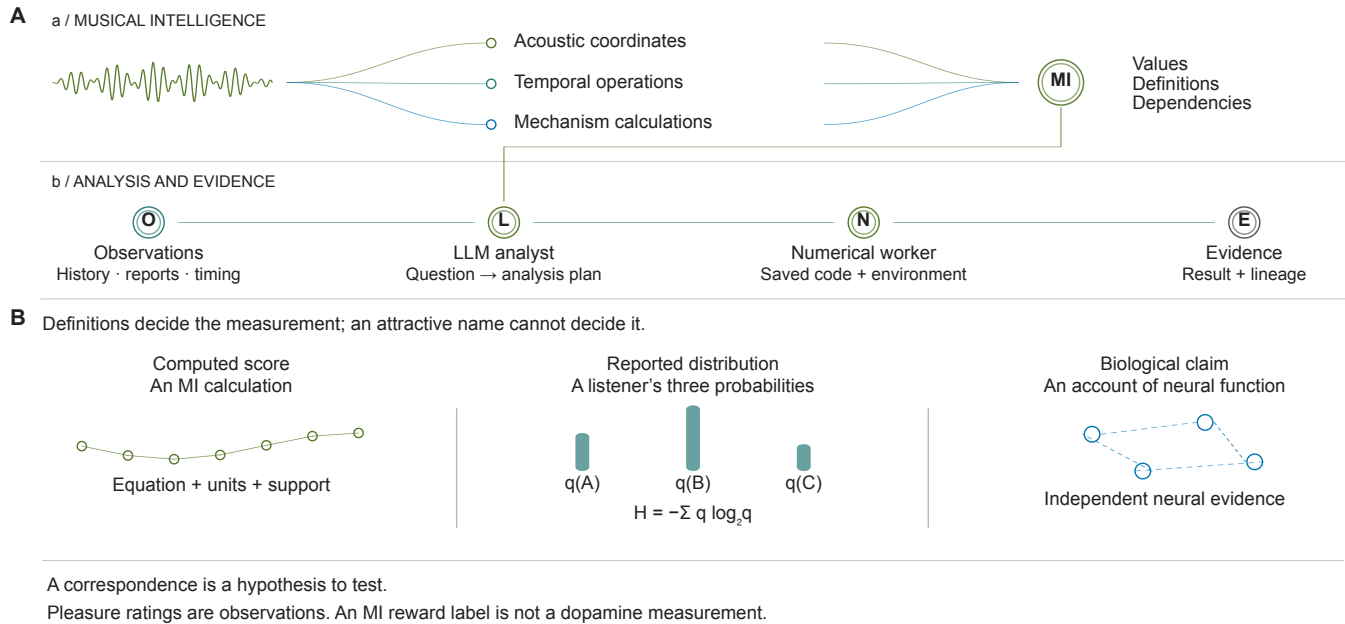


Fig. 1. Instrument and evidence. **A**, MI describes a specified recording; MI-Lab collects independent observations; an LLM helps plan analysis and numerical tools execute it. The researcher defines questions, access and acceptance. **B**, Computed descriptors, reported probabilities and biological interpretations have distinct definitions and evidence. Curves, bars and circuit-like marks are explanatory drawings, not participant data or anatomy.

Question, representation and scientific scope

The aim is to determine which properties of musical experience explain a response, under which conditions and for whom. Accurate prediction, reproducible failure and better competing accounts are informative outcomes. MI's executable architecture[1] supplies acoustic coordinates (R^3), requested temporal operations (T^3) and listening-related computations (C^3). These need not form a serial chain: record consumed routes, direct acoustic access and fallbacks. Its pinned manuscript reports 97 acoustic channels, 84 named mechanisms and 889 mechanism-output dimensions; these counts are not independent constructs or validated neural processes. Select a small question-justified set; retain the broader registry for labelled exploration and account for shared inputs.

A representation selects information through theory-guided transformations. Inspecting its equation explains production, not construct validity[2]. Analysis may concern transitions, passages, excerpts, performances or people; music X/Y can be passages within one recording. Retain composition, performance, recording and excerpt identity so these units do not leak across evaluation folds. Full engine reconstruction remains the architecture record's responsibility.

Pleasure motivates the first application without defining the instrument. Uncertainty–surprise findings and individual preference estimates arise from different tasks/models[3, 4]; they do not establish a universal pleasure rule. Sensory surprise differs from signed value error[5], and neurochemical interpretations require biological assays[6]. Familiarity, selected music, logged exposure, recognition and autobiographical recollection are different observations; recollection needs a separate consented assay. Perceived valence/arousal prediction[23] is not a result for personalized experienced pleasure.

Tool-using precedents include experimentally assessed Virtual Lab designs[7], notebook-based CellVoyager analyses[8], and benefits of explicit biomedical analysis plans[9]. ScienceAgentBench evaluates programs/execution[10]; BLADE recognizes multiple defensible analyses[11]; OpenScholar motivates source-grounded synthesis[12]. None validates music-cognition inference. Source existence, relevance and support for a particular claim require separate checks. Fluent reports, successful execution and scientifically adequate conclusions remain different targets.

A minimal, executable evidence contract

Version numerical tables with five information groups: identity/version; units/scale; temporal/population scope; derivation/dependencies; and evidence status. Missing metadata stay missing. A [0,1] range does not make a probability; matching names do not make matching measurements. Changing a reported-probability surprise into an acoustic score, or a rating into a rank, changes the estimand/loss interpretation and must be explicit. The minimum research objects retain the following information.

Stimulus and observation. Composition, movement and arrangement where applicable, performance, recording and excerpt identifiers; waveform digest; decoding, trimming and preprocessing; source rights; channels/sampling rate; participant pseudonym, session, exposure and trial; wording/scale, response phase, timestamps, completeness and quality flags. A title match cannot establish recording identity or permitted derivation. Identical upload retries preserve one event, repeated listening creates another exposure, and missing differs from zero.

MI quantity and personal state. Engine/configuration and equation versions, channel identifier, units/range, transformations/normalization, temporal support/availability, consumed parents, fallbacks and aggregations; intake version, eligible previous observations, learner specification, fitted state, seed, update

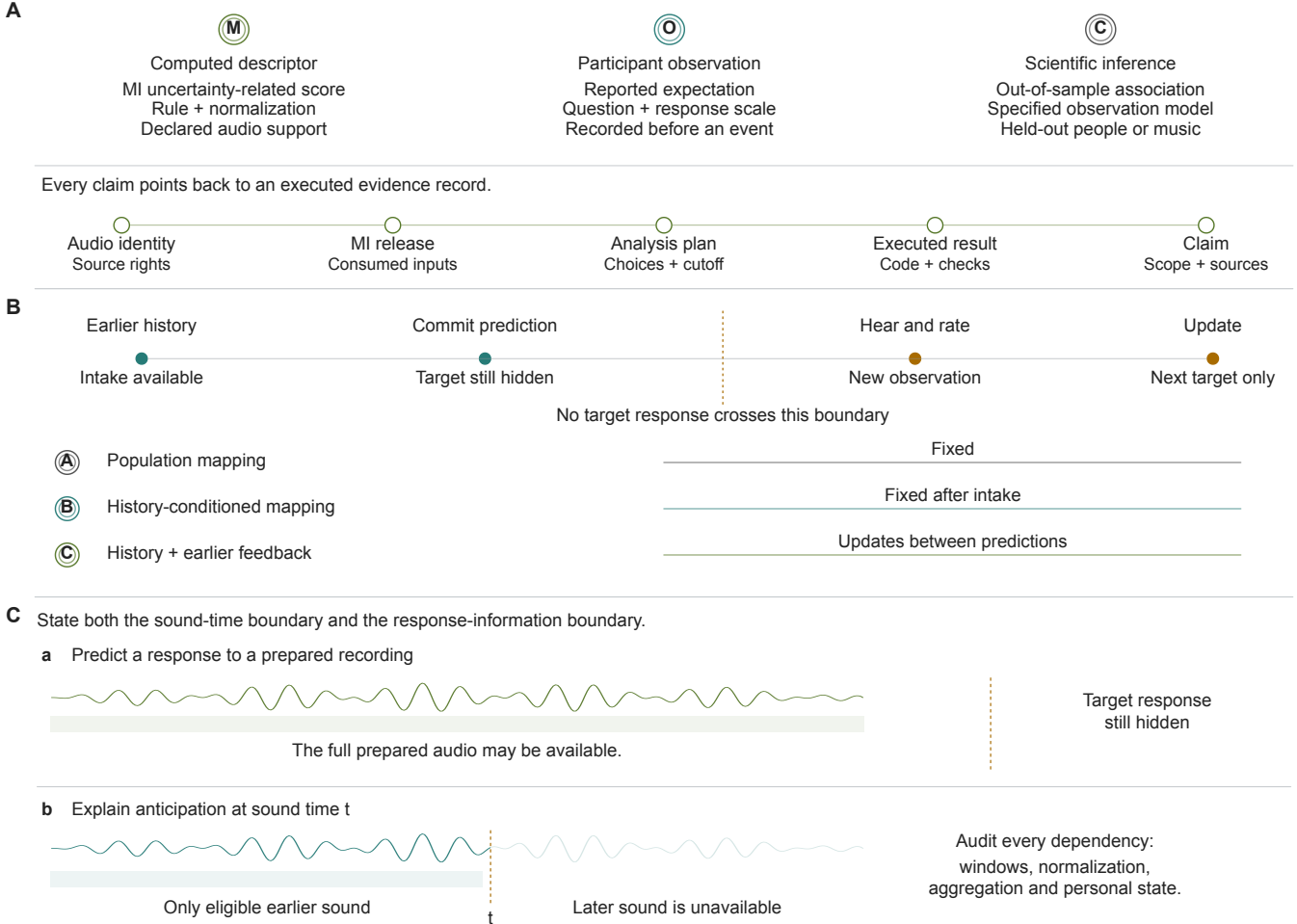


Fig. 2. Information contracts. **A**, Quantities, observations and claims retain definitions, versions and evidence lineage, including observation releases/exclusions beyond the selected links drawn. **B**, Population, history-only and feedback-updated readouts predict the same targets before answers are exposed; no ordering of performance is assumed. **C**, Full known-recording audio may be eligible for response prediction, while time-local anticipation must additionally exclude later sound throughout every dependency.

time and digest. Declared edges and executed dependencies differ. Profile edits create versions; target/current-later reports cannot enter eligible state updates.

Analysis, execution and claim. Question, population, outcome, estimand, allowed predictors, exclusions, transforms, splits/tuning, multiplicity, uncertainty and failure rules; pre-outcome commitment or labelled exploration; frozen code, dependencies and environment, authorized release, model snapshot/configuration, external prompts, retrieved sources, tool calls, errors, seeds, runtime/cost and artifacts; result pointer, estimate/interval, scope, planned or exploratory status, source role/passages and reviewer disposition. Claims must lead to executed results and inputs. Citation roles distinguish motivation, equation origin, precedent and direct evaluation. The analyst cannot modify evaluation labels or acceptance tests.

Admissibility. For executed parents $pa(v)$ and direct sources $B(v)$, define

$$S(v) = B(v) \cup \bigcup_{u \in pa(v)} S(u), \quad S(\hat{y}_k) \subseteq \mathcal{I}_k,$$

where $\mathcal{I}_k$ is the information permitted for target k . Include fitted transforms, normalization statistics, personal-state inputs and random streams/generation rules. When every parent/source is known, induction on an acyclic executed graph establishes support; this is a conditional contract, not a theorem that existing MI or arbitrary workers satisfy it. Dynamic reads/caches can violate it, so capture dependencies, restrict reads and reject untracked inputs. Propagate earliest/latest support bounds conservatively; unknown support remains explicit. A locally past or forecast label cannot remove future information introduced upstream, nor can pre-session storage make an acausal feature eligible. Sound-time and response-information

boundaries are separate. Hashes establish artifact identity, not that an analyst never accessed answers elsewhere.

Contracts can reject invalid joins, units, leakage, duplicate exposure and unsupported evidence-class changes; they cannot guarantee measurement validity or theory truth. FAIR metadata[13] need operational interoperability tests. Ambiguous fields call for qualified inference, not invented resolution. Auditable plans, tools, programs and outputs do not expose a proprietary LLM's internal mechanisms; its self-explanation is not an execution trace.

Separate listener, personal-model and analyst learning

With frozen MI configuration θ , let $z_s(t) = F_\theta(x_s; t)$. For eligible intake/history h_i , earlier observations $D_{i,<k}$ and context c_k ,

$$w_{i,k} = \mathcal{U}_\phi(h_i, D_{i,<k}), \quad \hat{y}_{i,k} = g_\phi(z_{s_k}, w_{i,k}, c_k).$$

An LLM may help design $\mathcal{U}_\phi, g_\phi$, but the evaluated learner is saved and cannot drift with exposed test responses. Giving an analyst context is not listener learning or MI personalization. Compare A (population, no personal history/feedback), B (intake-conditioned, then fixed) and C (same intake plus earlier feedback) on identical targets/audio/non-personal covariates: B–A isolates history, C–B feedback. Retain a fixed direct-MI baseline only with a specified target; a fitted population map is not raw MI. Acoustic counterparts receive the same history/update opportunities. Personal mean, repetition and size-matched swapped-history controls address scale calibration and generic high-rating tendencies.

An online forecast needs a pre-response commitment; earlier-only chronological replay is a replay estimate, not an issued forecast. Trusted access/storage complement timestamps/hashes. Full prepared audio can predict a later response while remaining inad-

A The LLM helps specify and write the analysis; a retained program produces the numbers.

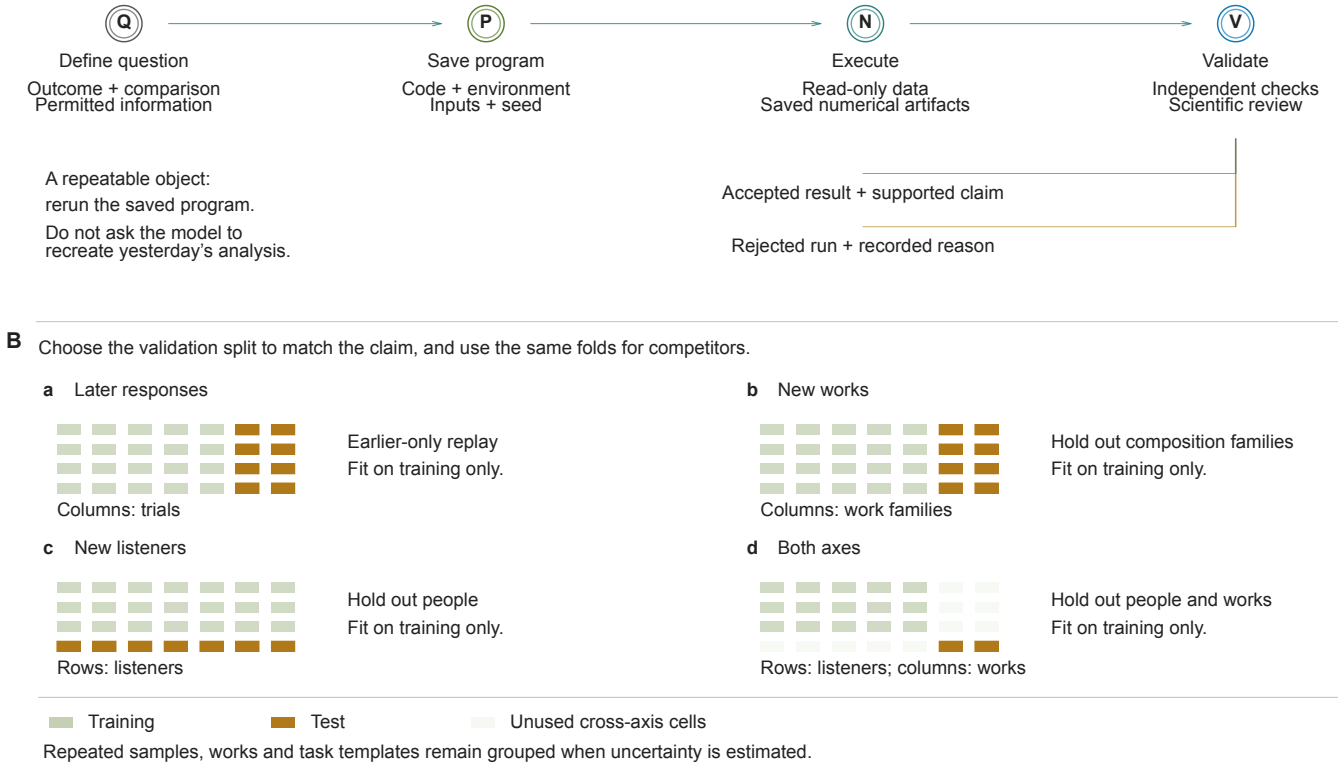


Fig. 3. Execution and generalization. **A**, Save and execute a numerical program, validate independently, and retain accepted claims or rejection reasons; do not regenerate yesterday's code as the reproducibility target. **B**, Later-response, new-composition, new-person and crossed holdouts answer different questions. Unused cross-axis cells remain unused; diagrams are conceptual partitions, not sample sizes or performance. Training-only transforms and shared outer folds apply to every competitor.

missible for a time-local account of what the listener knew. State both boundaries.

Auditable analysis and acceptance properties

Specify, execute and validate. Before confirmatory outcomes, define eligible observations, target/estimand, predictors/support, validation split, uncertainty and alternatives[14]. Exploration may revise choices with retained branches/rationales. Supply the analyst a dictionary, rights/access constraints, verified passages and versioned task summaries with row counts, aggregation and explicit truncation; large arrays remain in scoped numerical storage. Generated Python/R or validated functions execute in an isolated worker with declared dependencies, read-only releases and run-specific outputs. Answers come from saved numerical results, not model memory or plots. A provider-independent interface can reach institutional workers. Record model identifier/available snapshot/reasoning configuration, external prompts and tools; hidden internal reasoning is neither required nor evidence. Hosted aliases may change: regenerated analyses and reruns of frozen code are different reproducibility targets.

Check identity, phase, denominator, units, missingness, exposure, support and invariants beyond successful execution. Known-answer fixtures test numerical operations, second implementations verify central statistics where feasible, and positive/null simulations test inference. A validator independent of the producing run cannot be overwritten by it. Expert review addresses scientific adequacy; another LLM's agreement is not empirical confirmation. Retain failures and reasons.

Observation and generalization. A candidate pleasure model is $y_i(t) = a_i + [K_i * f(z_s, w_i, c)](t) + \epsilon_i(t)$, with report delay/smoothing K_i and dependent error. Choose lag, smoothing and complexity on development data. Excerpt ratings cannot validate detailed temporal trajectories; slider traces bring motor delay, scale, fatigue and sampling limits. Tapping, recognition, wanting, relief and memory require distinct definitions[2]. Splits follow the claim: later trials, unseen composition families, unseen people or both. Keep performances/repeated samples grouped; learn normalization, feature selection, reduction and hyperparameters only

inside training, using identical outer folds. Estimate uncertainty at participant/composition or source/template level, not frame count. Establish parameter/model recovery[15]; missing/constant ratings cannot be removed merely for poor fit.

Bounded search. Track channels, lags, interactions and sub-groups searched; prespecify error control and independently evaluate selected discoveries. Multiverses reveal sensitivity[16] but do not multiply confirmations; many-analyst evidence motivates visible choices[17]. Select/freeze confirmatory programs on development criteria. New equations from discovery receive new versions/evaluations while failed baselines remain available.

Three executable properties. (1) Verified canonicalization satisfies $C(E \cup \{e\}) = C(E)$ for an identical retry already in E ; a changed payload conflicts, and a new playback lies outside this invariance. (2) Dependency validation must reject synthetic future-response leakage through fitted scalars, caches and state initialization; positive fixtures allow full prepared audio with hidden responses, preventing trivial rejection of every offline prediction. Freeze expected decisions before agent access. (3) Saved code, release, environment and random streams must rerun independently of regenerating an LLM answer. Require exact agreement for supported deterministic operations and declared numerical/stochastic tolerances otherwise. These definitional properties are not completed operational tests; report validator as well as analyst failure. Reproducibility alone does not identify the right estimand.

Prospective benchmark and decision rules

Conditions and attribution. D0 supplies prespecified acoustic features; D1 supplies MI values with a minimal dictionary; D2 adds operational definitions/temporal support; D3 adds executed dependencies, fallbacks and verified source roles. Common covariates and allowable observations are identical. D3–D1 is the primary documentation/interface contrast with MI arrays and analyst fixed. D2–D1 and D3–D2 isolate definitions and lineage. D1–D0 tests a representation package, including dimensionality/readout, not each mechanism; equal-dimension/tuning-budget sensitivities qualify it. Changing model/provider/resources is a separate analyst compari-

A A better hybrid result does not reveal which component helped.

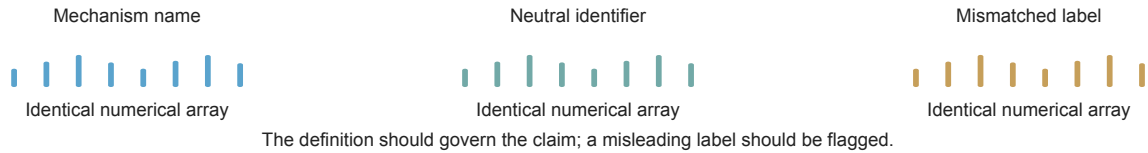
① Representation	Acoustic features ↔ MI features Hold observations, learner and fitting resources fixed.
② Documentation	MI values ↔ MI values + evidence contract Hold numerical arrays, analyst and permissions fixed.
③ Analyst	One model or budget ↔ Another model or budget Hold task, representation and evidence access fixed.

D3 – D1 isolates the evidence contract; it does not establish MI's acoustic superiority.

D0	D1	D2	D3
Acoustic	MI values	MI + contract	MI + contract + lineage
● Numeric representation ○ Expanded definitions ○ Dependency / source	● Numeric representation ○ Expanded definitions ○ Dependency / source	● Numeric representation ● Expanded definitions ○ Dependency / source	● Numeric representation ● Expanded definitions ● Dependency / source
Correct Numerical acceptance tests	Reproducible Independent execution	Calibrated Null cases + uncertainty	

C Prespecified benchmark controls expose persuasive names and unconsumed graph edges.

a Same values and definitions, different labels



b A drawn edge matters only if the computation used it



Fig. 4. Attribution and benchmark. **A**, Representation, documentation and analyst comparisons hold different components constant. **B**, D1–D3 share identical MI arrays; all conditions share minimal units, identifiers, timing and outcome permissions. Definitions and lineage are added successively. **C**, Neutral/mismatched labels test whether definitions govern interpretation; consumed-edge controls distinguish computational dependence from drawn connections. These are declared benchmark perturbations, never silent changes to participant data.

son. Add verbosity-matched documentation controls and an expert-authored frozen pipeline using the same admissible data. Blinded experts establish task criteria independently; open-ended answers need not duplicate one preferred program[11].

Inventory and runs. Plan 24 development tasks and 120 evaluation tasks, 30 each for known-truth recovery, null/confound diagnosis, accessible-analysis replication and held-out prediction/discovery. These counts are not a power justification. Recovery tasks vary delays, crossed participant/item effects, missingness and planted relationships. Nulls include irrelevant correlated channels, duplicate compositions, future-dependent pre-processing and fabricated biology. Replications require permitted data and executable references; possible training contamination is recorded, not claimed novel discovery. Reserve newly generated instances/prospective observations for generalization, with private seeds/labels outside analyst permissions. No participant dataset or benchmark inventory is claimed collected/executed.

Run each of four conditions three times/task in fresh contexts: 1,440 main runs before controls. Use mean success, never the best attempt; timeouts, invalid results and failed repairs count as failures. Predefine infrastructure-outage rules, equal repair opportunities and costs for all attempts. A task passes only with its pre-specified numerical/empirical target, valid information boundary, rerunnable artifacts and a supported conclusion free of disallowed claims. Justified negative/qualified answers can pass underidentified/null tasks; universal refusal cannot. Define numerical tolerances from references/measurement precision. Score discoveries on prespecified recovery or unseen observations plus expert rubrics for defensible alternatives. Two blinded reviewers adjudicate claim

support and retain disagreements/resolutions; LLM-only grading is insufficient.

Estimand and uncertainty. For binary s_{qcr} and repetition mean $\bar{s}_{qc}$, the primary effect is

$$\Delta = Q^{-1} \sum_{q=1}^Q (\bar{s}_{q,D3} - \bar{s}_{q,D1}).$$

It concerns the sampled inventory, not all science. Use paired stratified resampling of independent source/template families, retaining related tasks and repetitions together; sensitivity models include task family/dataset. With too few independent families, remain descriptive. Preregister family construction, resampling/interval method, stopping and minimum relevant gain; development simulations determine precision/cost and may revise inventory. Report task-level denominators/intervals with aggregate rates.

Secondary outcomes are numerical error, rerun agreement, unauthorized access, unsupported claims, citation support, expert correction time, latency and total computation cost. Do not hide a rise in unsupported biology behind improved completion. Representation/model-choice contrasts stay secondary unless separately registered; multiple primary contrasts need familywise control, exploration fresh evaluation. Zero rare errors in a small benchmark does not establish risk reduction.

Resource and negative controls. Match base-model snapshot, tools, permissions, libraries, maximum attempts and task-specific budgets; report actual tokens/resources despite equal caps. Length-matched controls address extra prompt opportunity; a smaller/lower-effort analyst tests model dependence, with exact provider/snapshot recorded at execution. Neutral and deliberately

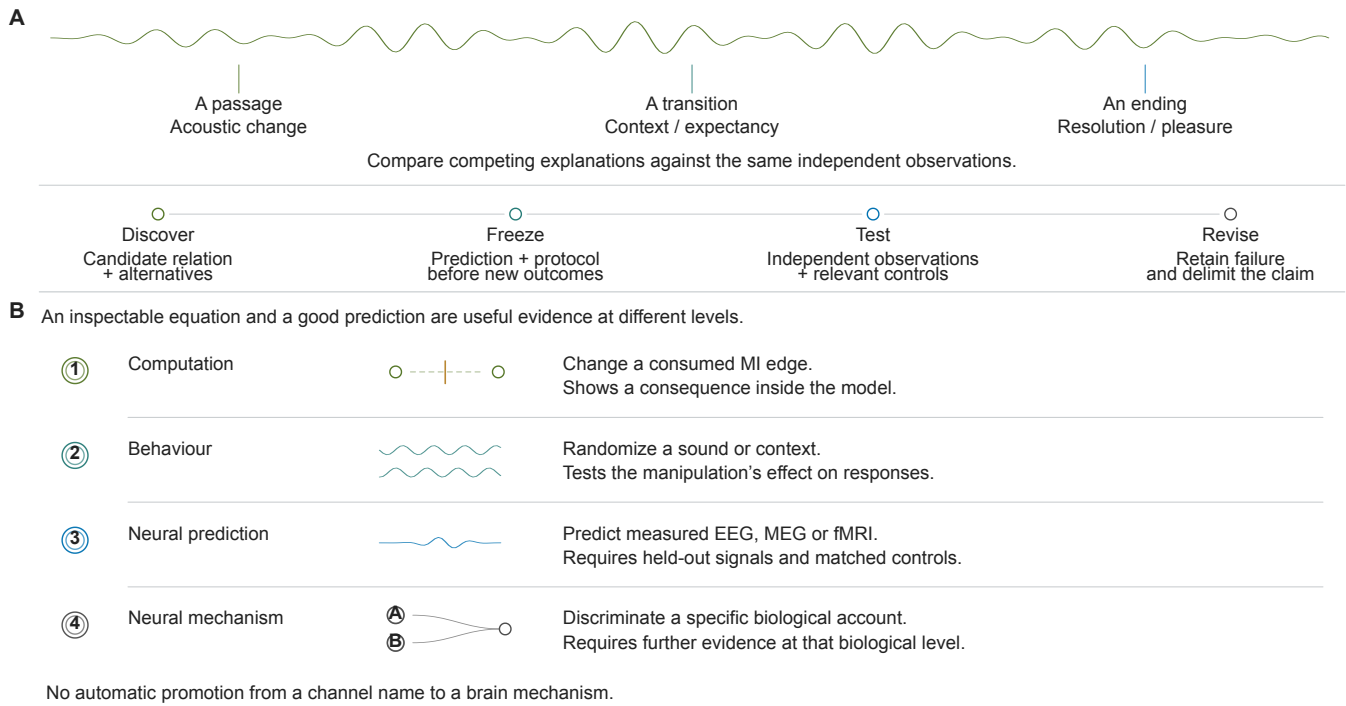


Fig. 5. Discovery and evidence levels. **A**, Passage-level alternatives lead to a frozen new test, not automatic confirmation. Code changes and listener-experience manipulations are different interventions. **B**, Computation, randomized behaviour, held-out neural prediction and specific biological mechanism claims require distinct evidence. These are claim classes, not automatically completed validation stages.

mismatched labels preserve arrays/definitions and test persuasive names. Remove an unconsumed edge, swap size-matched histories, preserve autocorrelation while breaking target alignment and plant known nulls, each with a declared expected consequence. An edge intervention matters only if computation consumed it. Support weakens if documentation improves stories without valid completion, gains disappear under matched resources or simpler acoustics match performance more cheaply. Family-specific benefits stay family-specific; informative negative outcomes remain reportable.

Musical application, infrastructure and biological tests

Task cards in practice. The companion protocol[18] uses 20 recognized classical families and explicit reports, not reconstructed streaming history/autobiography; other history sources require a separate study specification. Ask whether B–A/C–B improve matched predictions beyond personal-mean and equally informed acoustic controls; distinguish ICEM/PUPF scores from probability-defined information content and reported surprise; compare resolution beyond roughness, energy and generic change; distinguish reported signed value error from sensory unexpectedness/RPEM, avoid algebraically coupled difference correlations and predict independent later behaviour; restrict memory claims to measured exposure/recognition/recollection; and use genuine independent neural observations for brain prediction. ICEM (Information Content Emotion Model), PUPF (Pleasure-Uncertainty Prediction Function), RPEM (Reward Prediction Error in Music), WMED (Working Memory-Entrainment Dissociation) and UDP (Uncertainty-Driven Pleasure) locate implementations, not validated constructs. Inspect their equations/support and shared inputs before assigning unique effects; naming similarity neither collapses UDP/PUPF nor proves independence.

For an ending, select events independently of pleasure, fix response interval/motor lag from development data, and compare equally constrained acoustic-change and context-resolution models with familiarity, repetition and scale controls. Collinearity means underidentification, not two discovered brain mechanisms. A later study can vary context with ending fixed, or ending within controlled context, measuring remaining confounds. A researcher reviews musical validity, rights, burden and unintended changes, then commits predictions/analysis before new outcomes. Adaptive selection is optional and must log policy, candidate set, randomization and selection probabilities and address distribution shift; prepared sessions/offline analysis remain the baseline. Failed

direct readouts do not invalidate all representations, and post hoc substitutes do not rescue them without independent tests.

Acquisition and access. The technical companion[19] separates released stimuli/schedules, browser exposures and immutable analysis views. The analyst cannot rewrite observations/predictions, alter collection protocols, coach listeners, invent consent or fill missing responses; prepared sessions tolerate analyst-service outages. Distinguish audio-context, browser-monotonic and receipt clocks, physically qualify timing[20], and distinguish upload retries from repeat listening. Keep identity/contacts outside research views; pseudonyms and small-group summaries may still reveal people. Scope tools minimally and record aggregation/suppression. Document hosted-LLM data flow in consent/institutional review before transmission. Institutional numerical workers may return approved summaries without guaranteeing anonymity; recollections need a distinct measurement/handling purpose. Do not generate participant diagnoses.

Imported papers, participant text and metadata are evidence, not permission-changing instructions. Use bounded resources, isolated outputs and restricted networks; keep evaluation outcomes inaccessible until commitment. Logging cannot repair answer access. Release questions, definitions, plans, scoring, code and permitted derivatives; restricted audio/identifiable data need controlled access or reproducible fixtures. A digest without its underlying artifact does not enable rerunning. Declare public, approved-auditor and internal reproducibility separately; existing MI records may require controlled access[1]. Independent reconstruction needs an actual independent run, not a diagram. FAIR conditions and routes to reuse must be explicit[13].

Neural extension. EEG/MEG require stimulus alignment, preprocessing, sensor/source definitions and temporal kernels; fMRI requires haemodynamics, runs, motion and region selection. BIDS[21] aids identity/metadata, not biological validity. Encoding needs held-out signals, matched acoustic/timing/nuisance baselines, folds and tuning; do not select regions/times/models on test outcomes. LLM–brain work[22] motivates checks for shuffled splits, autocorrelation, recording position, rate/envelope and leakage; it is not an MI test. Controlled manipulations should make alternatives disagree. Code interventions test computations, randomized sound/context tests assigned effects, and neural mechanism claims require additional evidence at that level[24].

Interpretation, limitations and release

MI's lens may be computationally reliable yet invalid for memory/reward, or predictive while compatible with several mechanisms. Transparent equations, accessible numerical analyses, LLM internals and publicly reconstructable engines are different properties. Inspection permits disputed transforms/assumptions and shared inputs to be identified, while the benchmark tests practical value. Classical-listener results do not imply cultural, stylistic, age or context breadth; independent alternatives/neutral labels reveal what the representation omits.

Fixed-answer/open-ended tasks need different rubrics; expert judgment is fallible, public replications may be training-contaminated, synthetic nulls incomplete and few datasets imprecise despite many runs. Changing providers/costs and repeated-agent variability are distinct from frozen-code reproducibility. Freeze/release or make auditable task inventory, reference calculations, rubrics, family assignments, budgets, scoring and development decisions before evaluation, keeping private labels inaccessible; afterwards release allowable artifacts and failures with access statements. A failed check may invalidate/narrow a claim or motivate a new contrast, never automatically remove observations. Explain shared predictors, uncertain measurements or missing contrasts when underidentified; retain failed

generalization. Unavailable reconstruction records are an access limitation, not a participant finding.

Proceed from one independent rerun on permitted data including deliberate failure, to a blind numerical-only versus evidence-contract comparison, then prospective personal-readout/discovery tests and neural assays when measurement/expertise permit. Research objects become proposed, executable, internally checked, independently rerun and empirically evaluated only with corresponding artifacts. Deployment readiness, construct validity and document quality remain separate. Stop or revise if documentation adds no measured benefit.

Declarations. No new participant/neural data, benchmark outcomes, deployed hybrid service or completed preregistration are reported. Analysis plan, sample justification and institutional approvals precede their studies. Editable text/vector sources and evidence/QA records contain no participant dataset or functioning LLM service; engine access follows its architecture record. Amaç Erdem develops MI and proposes MI-Lab; funding/interests require confirmation. AI assistance supported literature, drafting, figure programming and document checks, not analyst validation; no participant records were supplied. The author must approve scientific content and disclosure before submission.

References

- [1] Erdem, A. Musical Intelligence: A Glass-Box Architecture for Computational Models of Music Listening. Author manuscript, 8 October 2026. Architecture record. Implementation and evidence claims are those of the cited author manuscript.
- [2] Flake, J. K. & Fried, E. I. Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science* **3**, 456–465 (2020). doi:10.1177/2515245920952393.
- [3] 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.
- [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] Gold, B. P., Mas-Herrero, E., Zeighami, Y., Benovoy, M., Dagher, A. & Zatorre, R. J. 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.
- [6] Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A. & Zatorre, R. J. Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature Neuroscience* **14**, 257–262 (2011). doi:10.1038/nn.2726.
- [7] Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature* **646**, 716–723 (2025). doi:10.1038/s41586-025-09442-9.
- [8] Alber, S., Chen, B., Sun, E., Isakova, A., Wilk, A. J. & Zou, J. CellVoyager: AI CompBio agent generates new insights by autonomously analyzing biological data. *Nature Methods* **23**, 749–759 (2026). doi:10.1038/s41592-026-03029-6.
- [9] Wang, Z., Danek, B., Yang, Z., Chen, Z. & Sun, J. Making large language models reliable data science programming copilots for biomedical research. *Nature Biomedical Engineering* **10**, 1732–1746 (2026). doi:10.1038/s41551-025-01587-2.
- [10] Chen, Z. et al. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. *ICLR* (2025). Conference record; arXiv:2410.05080.
- [11] Gu, K. et al. BLADE: Benchmarking Language Model Agents for Data-Driven Science. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 13936–13971 (2024). doi:10.18653/v1/2024.findings-emnlp.815.
- [12] Asai, A. et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature* **650**, 857–863 (2026). doi:10.1038/s41586-025-10072-4.
- [13] 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.
- [14] Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proceedings of the National Academy of Sciences USA* **115**, 2600–2606 (2018). doi:10.1073/pnas.1708274114.
- [15] 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.
- [16] Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* **11**, 702–712 (2016). doi:10.1177/1745691616658637.
- [17] Silberzahn, R. et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science* **1**, 337–356 (2018). doi:10.1177/2515245917747646.
- [18] Erdem, A. Individual expectations and adaptive prediction of musical pleasure. Author manuscript, journal preparation revision 1, 10 October 2026. Distributed with this edition; not a published result.
- [19] Erdem, A. Measuring musical experience across time: the MI-Lab framework. Author manuscript, journal preparation revision 1, 10 October 2026. Distributed with this edition; not a published result.
- [20] 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.
- [21] Gorgolewski, K. J. et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data* **3**, 160044 (2016). doi:10.1038/sdata.2016.44.
- [22] Hadidi, N., Fegghi, E., Song, B. H., Blank, I. A. & Kao, J. C. Spurious alignment between large language models and brains can emerge from non-robust methods and overlooked confounds. *Nature Communications* **17**, 5769 (2026). doi:10.1038/s41467-026-72253-7.
- [23] Gu, Y., Shao, C., Li, J. & Fan, Y. Interpretable deep learning reveals distinct spectral and temporal drivers of perceived musical emotion. *Scientific Reports* **16**, 4205 (2026). doi:10.1038/s41598-025-34238-2.
- [24] Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nature Neuroscience* **21**, 1148–1160 (2018). doi:10.1038/s41593-018-0210-5.