

Inspectable representations and auditable inference in computational music neuroscience

Amaç Erdem

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

Abstract

Computational models of music can generate richly labelled outputs without establishing that those outputs measure the corresponding cognitive or neural processes. This article proposes MI-Lab as an interface between an inspectable musical representation, independent observations and auditable analysis. Musical Intelligence supplies versioned stimulus descriptions; a personal readout can learn from earlier responses; a language-model analyst can help specify analyses that numerical programs execute. The framework separates these three processes and formalizes admissible information through the full support of consumed dependencies. It defines distinct tests for representation quality, documentation quality and analyst performance, rather than crediting every improvement to a single hybrid model. A proposed benchmark holds numerical arrays fixed while varying operational definitions and derivation records, with controls for persuasive labels, leakage, task dependence and unequal resources. Success requires a correct or appropriately qualified conclusion, valid information access and rerunnable numerical artifacts. The framework also specifies why prediction of a known-recording response differs from time-local anticipation, and why neural encoding differs from mechanistic explanation. No benchmark outcomes or participant results are reported. The central hypothesis is that explicit measurement and provenance contracts improve the validity of computational music research; the proposed comparisons permit that hypothesis to fail.

Methodological framework and testable benchmark proposal · No new participant, neural or benchmark results.

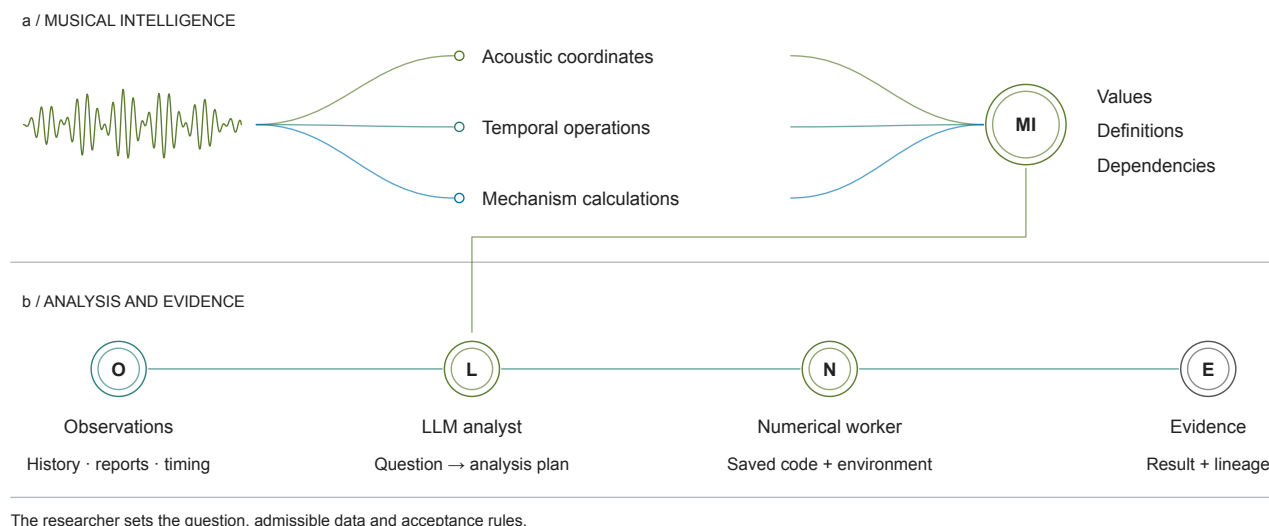


Fig. 1 | An instrument, an analyst and an evidence record. MI produces descriptors from a specified recording. MILab collects independent observations. A language model constructs an analysis plan, numerical tools execute it, and a retained record supports inspection and repetition. The analyst does not supply the participant's response. The waveform and traces are explanatory drawings.

Introduction

Which properties of an unfolding musical experience help explain a listener's response, under which conditions, and for whom? A useful research instrument should make this question easier to ask and harder to answer carelessly. A successful pleasure prediction is one possible finding. A reproducible failure, a dependence on familiarity, or a better competing account may be equally informative.

MI organises acoustic coordinates, requested temporal operations and listening-related calculations within an executable architecture[1]. Here, its outputs form a lens through which observations can be compared. The proposed analyst layer extends the use of that representation without making a language model the source of scientific truth. The test is whether the combination improves research decisions and reproducible findings.

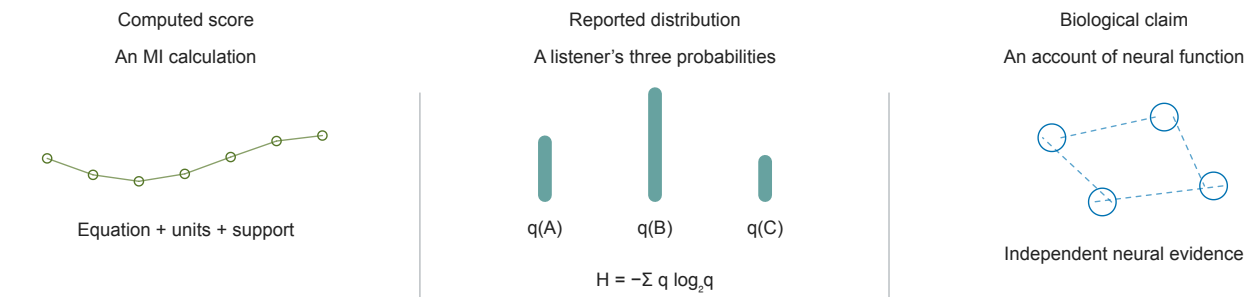
What the lens makes available

A lens selects and transforms information. It also imposes limits. MI is not a neutral inventory of every prop-

erty a listener might use: its feature definitions, temporal operations and mechanism equations embody modelling choices. Calling it an instrument therefore creates obligations to characterize those choices, quantify its useful range and expose failure. Inspection of an equation establishes how an output was produced. Construct validity requires a separate argument about what that output measures or predicts[2].

The pinned architecture manuscript describes a configuration with 97 acoustic channels, 84 named mechanisms and 889 mechanism-output dimensions[1]. These are author-reported implementation counts, not independent cognitive variables or validated neural processes. A study should select a small set justified by its question, while retaining the broader registry for documented exploratory analyses. Mechanisms that reuse inputs may be strongly dependent. A large repertoire increases the need for controlled model comparison rather than licensing an unrestricted search for significance.

Definitions decide the measurement; an attractive name cannot decide it.



A correspondence is a hypothesis to test.

Pleasure ratings are observations. An MI reward label is not a dopamine measurement.

Schematic marks show evidence classes, not values, a neural circuit or empirical construct validation.

Fig. 2 | A shared term does not make equivalent evidence. A computed MI descriptor, an elicited probability distribution and a biological interpretation have distinct definitions. Entropy is defined from the reported distribution, not inferred from an MI channel name. The bars, points and dotted circuit-like motif are abstract illustrations of evidence classes, not participant measurements or an anatomical circuit. Independent construct and neural evidence are required for the corresponding interpretations.

The lens can operate at several levels: a transition within one recording, a passage across performances, a full excerpt, or a comparison between listeners. Music X and music Y need not be different songs. They can be two passages with different computational descriptions. Musical identity must nevertheless be tracked at composition, performance, recording and excerpt levels so that these alternatives do not become accidental duplicates across evaluation folds.

MI's acoustic layer is designated R^3 , its requested temporal operations T^3 , and its listening-related calculations C^3 . These layers need not form a single serial chain: direct acoustic access and consumed upstream outputs depend on the implementation. MILab records the route actually used, including fallbacks, rather than treating a drawn connection as proof of computational influence. Full engine reconstruction is outside the present article; its architecture record remains the source for implementation-specific claims[1].

The scientific motivation is plural

Musical pleasure is a useful first application because plausible accounts make different claims about prediction and context. Cheung and colleagues reported an interaction between uncertainty and surprise in chord-related pleasure and neural responses[3]. Mas-Herrero and Marco-Pallarés subsequently found that the preferred level of surprise shifted with entropy in a melody-choice paradigm and related individual estimates to responses to real compositions[4]. These studies use different materials, observations and models. MILab should preserve such differences rather than convert them into one universal uncertainty-pleasure rule.

Reward-learning work also distinguishes an outcome's unexpectedness from whether it is better or worse than expected[5]. Neurochemical findings about musical emotion rely on methods designed to assess that level of biology[6]. An audio-derived reward or dopamine label is therefore a candidate computational interpretation, not a

measurement of reward circuitry or transmitter release. A pleasure rating does not remove that distinction.

Personal history adds another source of variation. Familiarity, preference, exposure and autobiographical meaning are related but non-equivalent. A list of selected music provides evidence about selection; a listening log provides recorded exposure under a platform's logging rules. Neither is a direct observation of autobiographical memory. Event-specific recollection requires a suitable report or assay and a separate consented research question.

Why an LLM-assisted analyst

The proposed division of labour has precedents, but its domain-specific value remains untested. The Virtual Lab combined language-model agents with specialist computational tools and assessed designed nanobodies experimentally[7]. CellVoyager developed and executed single-cell analyses in notebooks, evaluated against published studies and expert-reviewed cases[8]. These studies motivate a tool-using analyst; they do not establish that music-cognition inferences will be correct.

The engineering lesson is that analysis structure matters. Wang and colleagues found that an explicitly developed analysis plan improved performance on their biomedical programming benchmark[9]. ScienceAgent-Bench evaluates generated programs and their execution, while BLADE addresses the multiplicity of defensible analysis decisions[10, 11]. A fluent report, successful program execution and a scientifically adequate answer are different evaluation targets.

Literature support is another task. OpenScholar combines retrieval and verification to improve source-grounded synthesis[12]. MILab would use source retrieval to establish what an article actually studied, which measure it used and how it supports a proposed interpretation. The source's existence, its relevance and its support for a specific claim are checked separately.

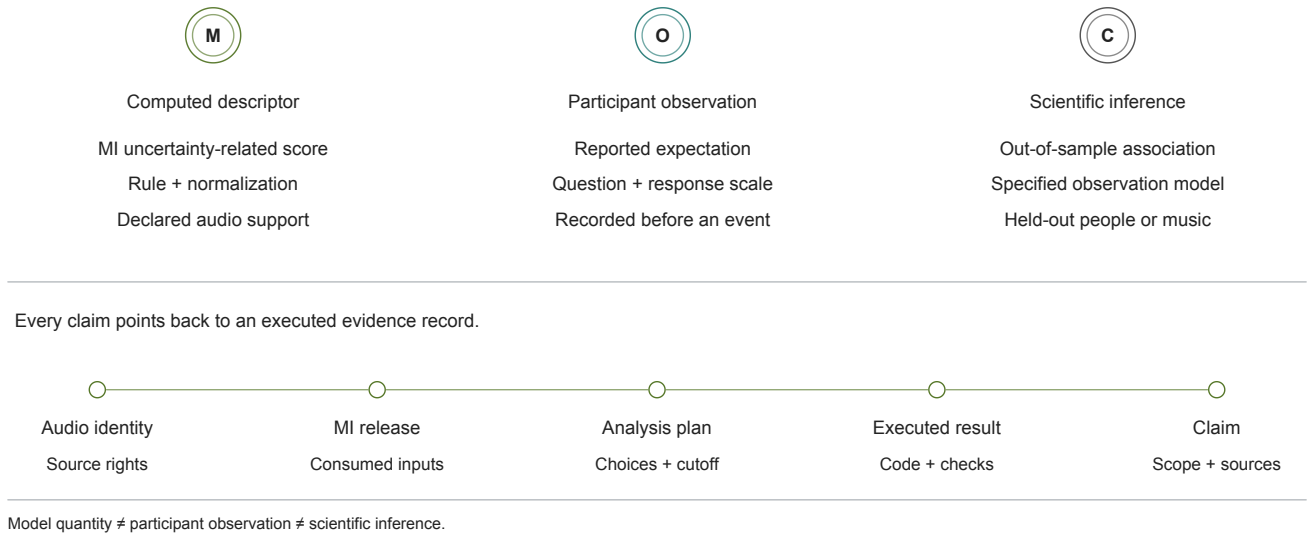


Fig. 3 | Keep a computed quantity, an observation and a claim distinct. Definitions and identifiers attach to values before analysis. A computational uncertainty score is not silently converted into a reported probability or information-theoretic entropy. The lower sequence shows selected lineage links; the complete record also references observation releases, exclusions and source passages. Provenance supports inspection, not automatic validity.

The proposed contribution is thus a domain-specific interface between inspectable music representations, human observations and an auditable research workflow. It must earn its value against simpler features, ordinary documented analyses and other representation choices.

The evidence contract

Each accessible quantity has a stable identifier, numerical representation and interpretation record. The proposed contract contains five groups of fields: identity and version; units and scale; temporal and population scope; derivation and dependencies; and evidence status. The analytical table and this contract are versioned together. Missing fields are reported as missing rather than inferred from an attractive variable name.

Identity specifies the recording digest, composition family, performance, excerpt, participant pseudonym and exposure where applicable. Scale specifies units, admissible range, transformations and whether a displayed value has been normalized. A value in $[0, 1]$ is not a probability merely because its range resembles one. Two similarly named channels may use different transformations and must retain separate identifiers.

Temporal scope specifies which samples or events enter a calculation, when the result becomes available, and which alignment rule links it to an observation. The full dependency support matters. Recording-wide normalization can incorporate later audio even if a local window is labelled past. A centred operation can use samples on both sides of an event. A channel labelled forecast is not a demonstrated forecast without an admissible information boundary and prospective evaluation.

Derivation records the engine release, request, equation version, consumed inputs, fallback flags and any later aggregation. It distinguishes declared from executed dependencies. Evidence status separates a computed descriptor, reported observation, fitted prediction, exploratory association, confirmed comparison and biological interpretation. A source citation is attached with its role: motivation, equation origin, implementation precedent or direct evaluation.

Formalizing admissible information

The methodological contribution is an explicit interface between representation, observation and inference. It can be stated independently of a particular language model. Let a computational graph contain values v , executed parent relations $pa(v)$ and direct source observations $B(v)$. Define the support of a value recursively by

$$S(v) = B(v) \cup \bigcup_{u \in pa(v)} S(u). \quad (1)$$

Support includes fitted transformations, participant-state inputs and normalization statistics, not only the samples in the final local operation. Given the information set $\mathcal{I}_k$ permitted before target k , a prediction is admissible only if

$$S(\hat{y}_k) \subseteq \mathcal{I}_k. \quad (2)$$

For stochastic computation, the random seed or stream and its generation rule are also recorded. Admissibility constrains what can enter a prediction; it says nothing by itself about its accuracy or psychological validity.

This recursive definition has a useful consequence. If each direct source and fitted parameter has verified support, and each executed operation reports all of its parents, induction over an acyclic execution graph establishes the support of its outputs. A nominally past-centred final operation cannot remove future information already introduced by recording-wide normalization. Conversely, full-stimulus features are admissible for a known-recording response-prediction task when the response itself is withheld. The sound-time and response-information boundaries must therefore be specified separately.

These are requirements of a proposed execution contract, not a theorem that the current MI engine or an arbitrary worker already satisfies them. Dynamic code, unreported file reads and shared caches can violate the premises. The implementation must capture consumed dependencies, restrict external reads and reject untracked inputs. A hash establishes artifact identity; it does not establish that the analyst never saw a target through another route.

Measurement types and transformation validity

A value record is typed by measurement scale, support and evidential role. An information-theoretic surprise derived from a reported probability distribution cannot be substituted by a normalized acoustic score merely because both are named surprise. The substitution would change the estimand. Likewise, transforming a rating into a rank changes which errors a loss function penalizes. A valid analysis must declare the transformation and interpret its target accordingly.

The contract is deliberately smaller than a complete account of scientific meaning. It provides checks that can reject concrete category errors: inconsistent units, impossible joins, target leakage, exposure duplication and unsupported changes of evidence class. It cannot guarantee that a construct is well measured or a theory is true. Those questions require task validity, competing models and independent observations. This separation follows the broader aim of cognitive computational neuroscience: models should connect task performance and measured signals through testable assumptions[24].

From the contract to a scientific claim

A claim record identifies the question, population, outcome, estimand, data release, analysis plan, code, result and uncertainty. It includes scope restrictions and whether the analysis was planned or exploratory. The system should allow a reader to move from a sentence to the numerical artifact and then to the relevant inputs. A citation to a general theory cannot substitute for that chain.

The contract is motivated by machine-actionable stewardship[13], but richer metadata alone does not guarantee interoperability. Units, identity, support and evidence types require operational tests across releases. Ambiguous fields are a reason to qualify an analysis, not a reason for the LLM to invent a resolution.

Different forms of transparency

MI can expose its selected computations. The analyst can expose its external plan, requested tools, executed programs, decisions and outputs. These properties do not expose the internal mechanisms of a proprietary LLM. MI-Lab therefore claims auditable analysis, not an entirely interpretable hybrid model. Explanatory prose or a model's account of its reasoning is not an execution trace. The reproducible object is the retained analysis program with its inputs and environment; it should not require asking the same LLM to write the program again.

Separate three learning processes

The listener can learn from sound, context and repeated exposure. A personal statistical model can learn a mapping from descriptors to that listener's responses. The research analyst can acquire more information about a dataset or develop a better analysis. These processes have different states, evidence and tests. Sending an additional table to an LLM is not, by itself, evidence that MI has learned a listener.

Let x_s be recording s and let θ identify a frozen MI configuration. Its representation is

$$z_s(t) = F_\theta(x_s; t). \quad (3)$$

The contract defines the actual audio support of F_θ . Let h_i denote eligible history for listener i , and $D_{i,<k}$ the observations available before target k . A personal state and

response prediction are

$$w_{i,k} = \mathcal{U}_\phi(h_i, D_{i,<k}), \quad (4)$$

$$\hat{y}_{i,k} = g_\phi(z_{s_k}, w_{i,k}, c_k), \quad (5)$$

where c_k contains permitted context and ϕ fixes the learner specification. The LLM may help design $\mathcal{U}_\phi$ and g_ϕ , but an evaluated version is saved and executed explicitly. Its definition is not allowed to drift as test responses become visible.

Three matched prediction conditions

A population readout uses no personal history or feedback. A history-conditioned readout uses only the intake state, fixed for evaluation. A feedback-updated readout starts from the same intake state and incorporates earlier eligible responses. The contrasts are B versus A for history and C versus B for feedback. All candidates receive the same target audio and common non-personal covariates, and are scored on the same target responses.

An existing fixed direct MI readout is retained as an additional baseline when its target interpretation is specified. A fitted population mapping must not be relabelled as raw MI.

A corresponding acoustic-only family receives the same personal information and update opportunities. Additional controls include a personal mean, simple repetition effects and a swapped-history condition matched for history size. These distinguish structured preference learning from response-scale calibration or merely learning that a participant tends to give high ratings.

What counts as a future prediction

For an online claim, the prediction and its information cutoff must be committed before the target response is collected. A retrospectively fitted program can also be evaluated by chronological replay, provided that every replay step hides the current and later responses. Such a result is a replay estimate, not a previously issued forecast. A timestamp or hash establishes which artifact was committed; trusted storage and access control are still required to establish what information was available.

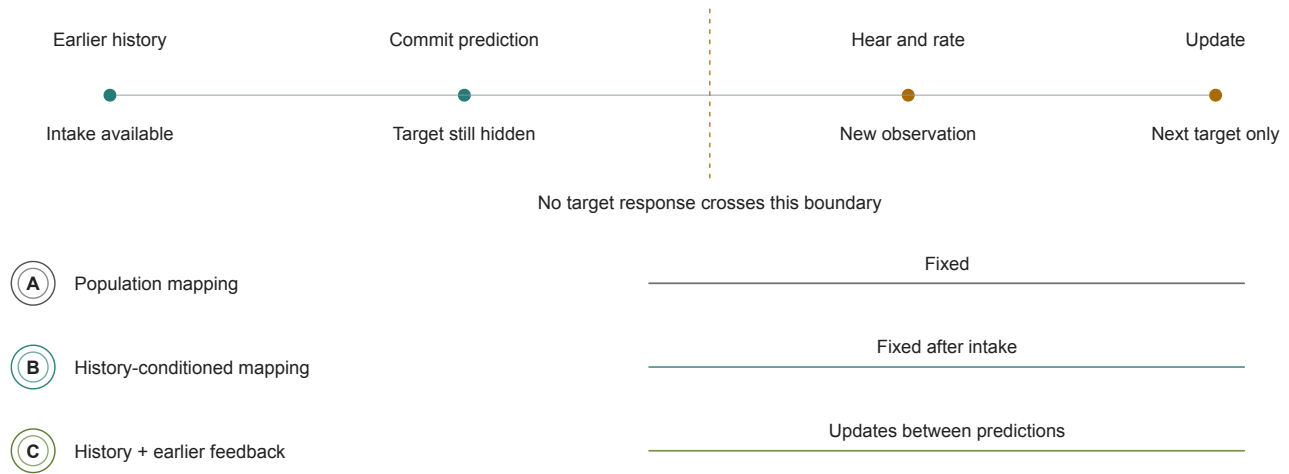
Full target audio may be available before playback for an offline stimulus-based prediction. This must not be confused with a causal, time-local account of what the listener knew at second t . For a moment-by-moment prospective account, the feature support must obey the corresponding sound-time boundary. Both tasks are useful, but they answer different questions.

An auditable analysis cycle

Translate the question before inspecting outcomes

The investigator states a question and an intended claim type. The analyst first constructs an analysis specification: eligible observations, target, estimand, permitted predictors, time support, validation split, uncertainty method and competing explanations. For a confirmatory test these choices are frozen before access to evaluation outcomes. Exploratory work permits revision, but every branch and its rationale remains recorded. This follows the distinction between prediction and post-outcome explanation that motivates preregistration[14].

The LLM receives a data dictionary, rights and access constraints, approved source passages and task-relevant summaries. Large arrays remain in numerical storage and

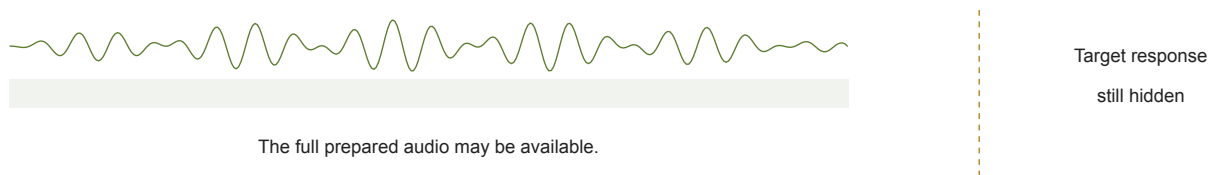


Listener learning $\neq$ personal-model updating $\neq$ the analyst acquiring context.

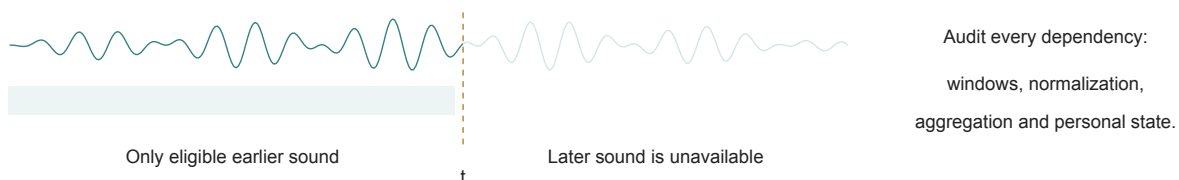
Fig. 4 | Learning is evaluated across an explicit information boundary. All candidate readouts predict the same target before its response is exposed. A uses a population mapping, B adds an intake-conditioned personal state, and C additionally updates from earlier eligible feedback. These comparisons concern a proposed personal readout, not an undocumented change to the frozen ML engine. No performance ordering is assumed.

State both the sound-time boundary and the response-information boundary.

a Predict a response to a prepared recording



b Explain anticipation at sound time t



A "past" or "forecast" label is insufficient if an upstream computation consumed future information.

Fig. 5 | Two useful prediction tasks require different information boundaries. An offline prediction of response to a known recording can use its full prepared audio while withholding the response. A time-local anticipation claim additionally excludes later sound and every dependency that incorporates it. Upstream normalization, aggregation, temporal operations and personal state all require eligibility checks. The waveform is schematic. Neither a past-window label nor a forecast channel establishes compliance by itself.

are accessed through scoped tools. This avoids treating a long prompt as a numerical database. Tables returned to the model are versioned, with row counts and aggregation rules; truncation is explicit.

Execute the specification

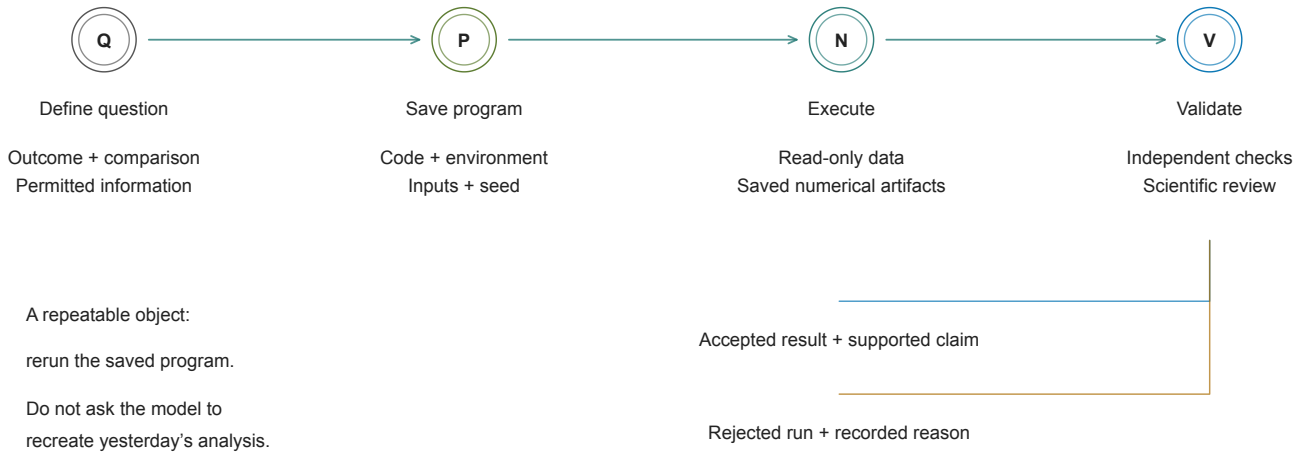
The analyst may generate Python or R, call validated analysis functions, or combine both. Code executes in an isolated worker with declared dependencies and a read-only data release. New artifacts are written to a run-specific location. Numerical answers must be taken from executed outputs, not generated from memory or inferred from a plot. A scoped tool interface can connect the analyst to an institution-controlled numerical worker; the scientific contract does not depend on a provider-specific execution product.

The execution record contains code and environment digests, model identifier, available model snapshot, reasoning configuration, external prompts, retrieved sources, tool calls, random seeds, errors and output hashes. Internal hidden model reasoning is neither required nor treated as evidence. Hosted-model aliases can change, so replaying the analyst's generation is a different reproducibility target from rerunning its frozen code.

Validate beyond successful execution

A program can finish and still use the wrong participant join, outcome phase, denominator or time alignment. Checks therefore include row-level identity, units, missingness, duplicate exposures, temporal support and expected invariants. Known-answer fixtures validate se-

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



Successful execution is necessary but insufficient: identity, timing, numerical and claim checks must also pass.

Fig. 6 | An auditable analysis connects a question to executed evidence. The LLM may help construct the specification and numerical program. A saved environment and read-only input release produce numerical artifacts; independent checks and scientific review determine whether the resulting claim is supported. Rejected runs retain their reason. Reproducibility concerns re-executing the saved program, not asking a mutable model to generate an identical program or treating generated prose as a calculation.

lected numerical operations. A second implementation verifies central statistics where feasible; synthetic positive and null controls test the complete inferential route. These checks are part of the method, not a claim that this manuscript has already executed the research workflow.

A validator is independent of the code-producing run and cannot be overwritten by it. Expert review addresses scientific adequacy that unit tests cannot establish. A second LLM can assist review, but agreement between models is not independent empirical confirmation. A rejected analysis remains in the record with its failure reason.

Model the observation, not just the feature

For continuous pleasure, a possible observation model is

$$y_i(t) = a_i + [K_i * f(z_s, w_i, c)](t) + \epsilon_i(t), \quad (6)$$

where K_i represents response delay and smoothing, and ϵ_i can be temporally dependent. This is a proposed statistical family, not an MI equation. Lag estimation, smoothing and model complexity are chosen using development data. Retrospectively maximizing correlation over many lags on test recordings would change the test.

The endpoint must match the claim. An excerpt rating supports an excerpt-level analysis; it cannot validate a detailed within-excerpt trajectory. A slider trace permits temporal questions but brings motor delay, scale use, fatigue and sampling limitations. Additional tapping, recognition, wanting, relief or memory questions require their own operational definitions. Keeping them distinct is essential to measurement transparency[2].

Preserve the unit of generalization

Adjacent samples, repeated excerpts and performances of one composition are not independent observations. The split follows the proposed use: later trials for within-person learning; held-out composition families for new-music prediction; held-out people for population transfer; or both axes for broader generalization. Training-

only preprocessing includes normalization learned across items, feature selection, dimension reduction and hyper-parameter tuning. Every comparison uses the same outer folds.

Uncertainty is reported at the appropriate sampling unit, for example through participant and composition effects or a prespecified resampling procedure that respects both. Sample count at the audio-frame level does not establish the number of independent people or musical items. Model and parameter recovery should precede strong interpretations of fitted personal parameters[15]. Missing responses and constant ratings are retained or excluded by explicit rules, never because they reduce a desired correlation.

Bound the search

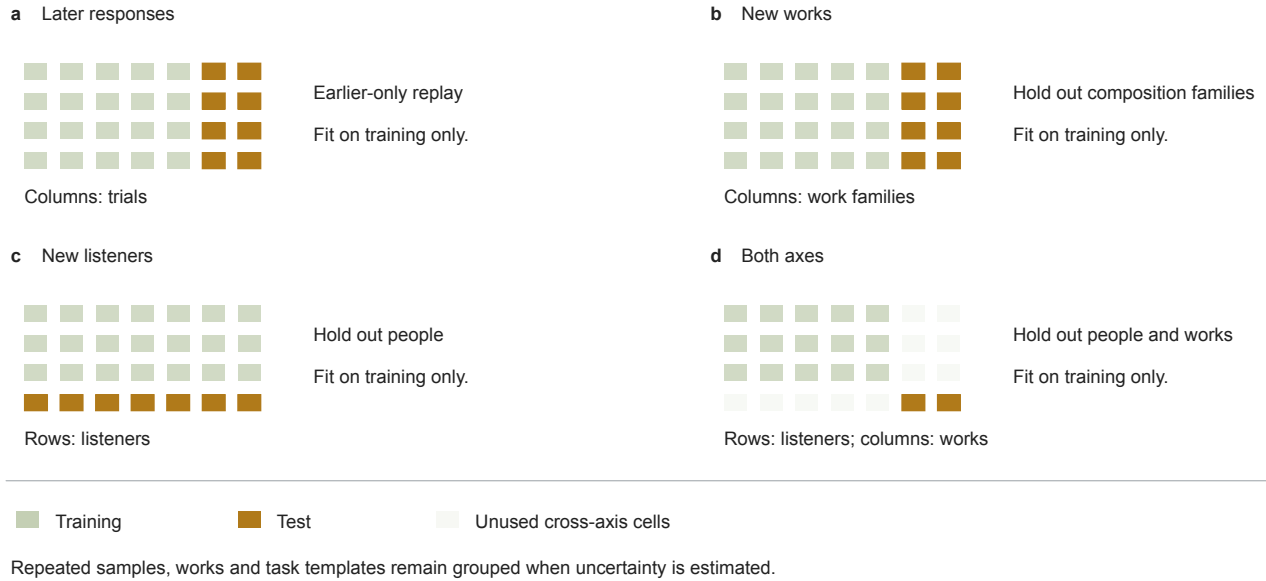
Exploration across channels, lags, interactions and subgroups is a search over many hypotheses. MILab records its breadth, uses prespecified error control where appropriate, and requires fresh evaluation of selected discoveries. A multiverse can expose sensitivity to defensible choices[16]; it does not turn all searched analyses into independent confirmations. The many-analysts literature demonstrates why seemingly reasonable decisions deserve this visibility[17].

For confirmatory work, the best generated program is selected on development criteria and then frozen. For discovery, the LLM may propose new mechanisms or questions, but any changed MI equation receives a new version and a separate evaluation. The baseline implementation remains available even when it performs poorly.

Analytical properties and executable acceptance targets

Three properties make the framework testable before participant recruitment. First, duplicate delivery of an unchanged event must leave the statistical dataset invariant. If $C(E)$ canonicalizes event records by verified identity and payload, then for an identical retry e already rep-

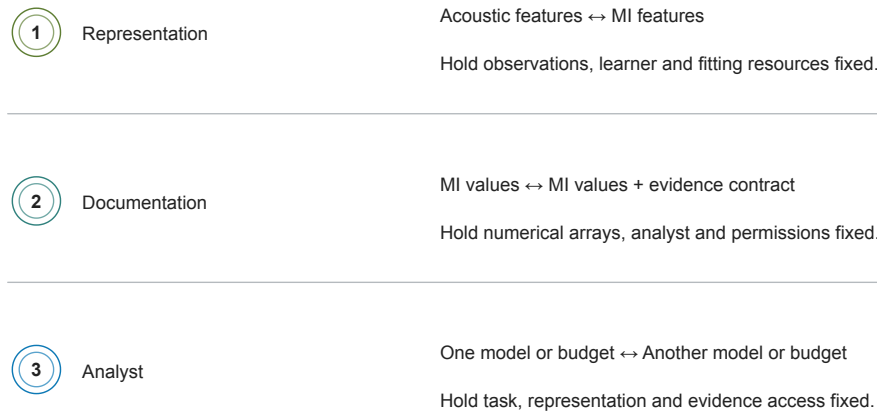
Choose the validation split to match the claim, and use the same folds for competitors.



Illustrative fold patterns. A dense trace is not a large sample of independent people.

Fig. 7 | The withheld unit must match the intended generalization. The panels illustrate later-response, new-work, new-listener and crossed generalization. Coloured cells indicate conceptual training and test partitions, not the study sample size or a measured score. For the both-axis example, unused cross-axis cells are faded. All transformations are learned within training, competing models share outer folds, and uncertainty respects participant, composition or source/template dependence appropriate to the analysis.

A better hybrid result does not reveal which component helped.



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

Comparison design, not benchmark results · A valid negative or qualified answer can be a success.

Fig. 8 | Representation, documentation and analyst effects are different estimands. The primary documentation comparison keeps MI numerical arrays and analyst configuration fixed. A representation comparison requires matched information and fitting resources; an analyst comparison holds the scientific task and input representation fixed. No component is credited with an effect demonstrated only for the combined system. The plate specifies tests, not observed performance.

resented in E , $C(E \cup \{e\}) = C(E)$. A changed payload under the same identifier is an error rather than an extra sample. A second playback has a new exposure identity and is therefore outside this invariance. This is a data-construction property, not a psychological assumption.

Second, admissibility must be closed over the full consumed dependency graph. A synthetic fixture will deliberately introduce future-response data through a fitted scaler, a cache and a state initializer. A correct validator rejects each route even if the final readout uses an apparently eligible feature table. Independent fixtures preserve full prepared audio while hiding responses, so that a validator cannot pass simply by rejecting every useful offline

prediction. Expected decisions are specified before agent access.

Third, a numerical result must be reproducible independently of regenerating the analysis with an LLM. Fixing the saved program, input release, numerical environment and random streams defines the computational object to repeat. Deterministic operations should match exactly where the environment supports that requirement; floating-point or stochastic operations require declared tolerances and replication criteria. Reproducibility of the saved computation does not imply that a future hosted model will produce the same program or that the program estimates the right quantity.

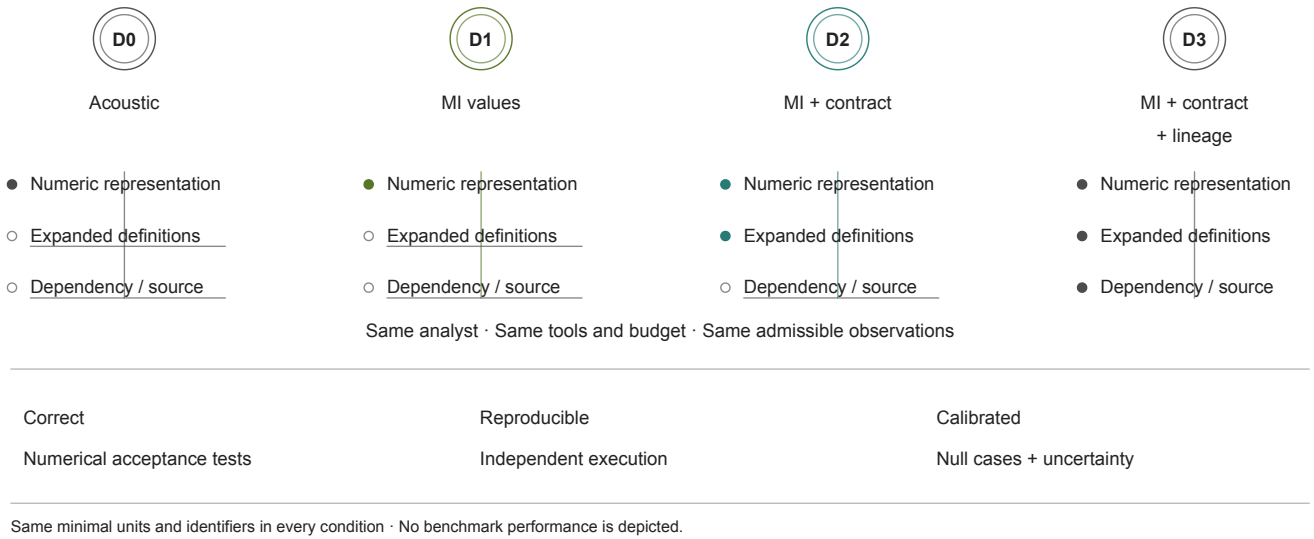


Fig. 9 | Separate useful features from useful documentation. All agent conditions receive valid numerical units, identifiers, timing conventions and the same outcome access policy. The MI conditions use identical numerical arrays. Rich operational definitions and then derivation/source lineage are added in successive conditions. These contrasts estimate different contributions; the diagram reports no benchmark outcomes.

The algebraic properties above follow from their stated definitions. Their operational realization remains to be tested. They are not empirical evidence of agent accuracy, brain correspondence or improved musical-pleasure prediction. The benchmark must report validator failures as well as analyst failures, because a brittle acceptance rule can make a misleading system appear reliable.

Locating the source of an apparent gain

The full hybrid system has at least three independently variable components: the numerical representation, its scientific documentation and the analyst. The main D3–D1 contrast holds the MI arrays fixed and varies the evidence contract. It therefore estimates a documentation/interface contribution, not MI’s superiority to acoustics. The acoustic-versus-MI comparison targets representation under matched information and readout resources. Changing model provider or reasoning budget targets the analyst and belongs to a separately defined comparison.

This separation avoids crediting every positive result to the architecture. A well-documented acoustic baseline may support more valid analyses than a poorly specified rich representation. A stronger analyst may improve both conditions without benefiting from MI-specific semantics. Conversely, useful documentation may reduce unsupported biological claims even when numerical error is unchanged. The success criterion must expose these cases rather than combine them into a promotional score.

Current interpretable music-emotion prediction offers an important comparison class[23], but performance on perceived valence and arousal is not a baseline result for personalized experienced pleasure. Model–brain agreement also requires specific methodological controls: recent analyses show how temporal and analytical confounds can create apparent alignment[22]. MI-Lab consequently treats generalization units, preprocessing support and negative controls as part of the question being tested, not as an afterthought to a high correlation.

A prospective evaluation

The central methods question is whether an inspectable representation improves the accuracy and reproducibility of LLM-assisted research. Four agent conditions are

proposed. D0 provides a prespecified acoustic representation; D1 provides MI values with a minimal data dictionary; D2 adds operational definitions and temporal support; D3 adds executed dependency records, fallback information and verified source roles. Common participant covariates and allowable observations are identical.

D1 versus D0 evaluates a representation package, including its dimensionality and chosen readout. It cannot by itself identify the benefit of every mechanism. D2 versus D1 isolates the added operational description, while D3 versus D2 isolates the additional lineage and source information. Equal-dimension and equal-tuning-budget sensitivity analyses qualify representation comparisons. A verbosity-matched documentation control tests whether an apparent benefit merely reflects a larger prompt.

An expert-authored, frozen numerical pipeline provides a separate reference using the same admissible data. Experts define and independently review task criteria before seeing agent-condition identities. For open-ended questions, several analyses may be valid; scoring cannot require an exact copy of one expert’s preferred code. This limitation motivates both task-specific numerical checks and explicit scientific rubrics[11].

A proposed task inventory

The initial specification contains 24 development tasks and 120 evaluation tasks, with 30 in each of four strata: known-truth numerical recovery, null/confound diagnosis, replication of an accessible analysis, and held-out prediction or discovery. These counts define a planning inventory, not a demonstrated adequate sample size. The final preregistration will revise them if development estimates show inadequate precision or cost.

Known-truth tasks include delayed responses, crossed participant/item effects, controlled missingness and planted feature relationships. Null tasks include correlated but irrelevant channels, duplicate compositions, future-dependent preprocessing and a fabricated biological interpretation. Replication tasks require rights-compatible data and an executable reference. Published analyses are used to assess reproduction, not as proof of novel discovery; their presence in model training is documented as a possible advantage.

Newly generated evaluation instances and prospective observations are reserved for generalization. The evaluator holds private seeds and labels outside the analyst’s permissions. Public benchmark instructions may describe the task class, but cannot reveal private outcomes. No participant dataset or benchmark task inventory has been collected or executed for this manuscript.

Evaluation estimands and analysis plan

Primary contrast and outcome

The primary comparison is D3 versus D1: does a full evidence contract improve valid task completion when the numerical MI representation is unchanged? A task passes only if it meets its prespecified numerical or empirical target, respects the allowed information boundary, produces rerunnable artifacts and supports its main conclusion without a disallowed claim. For a null or underidentified case, a justified negative or qualified answer can be the correct result. Merely refusing every difficult task cannot score as general success.

Numerical tolerances are defined per task from reference calculations and measurement precision. Discovery tasks are scored on recovery of a prespecified relation or performance on unseen observations, with a separate expert rubric for alternative defensible findings. Generated prose is not evaluated solely by another language model. Two blinded reviewers adjudicate claim support; disagreement and its resolution are retained.

Each condition is run three times per evaluation task in fresh contexts, yielding 1,440 planned main runs before controls. The reported task-level success is the mean across all three runs, not the best attempt. Timeouts, invalid outputs and unsuccessful repairs count as failures; infrastructure outages are separately labelled using rules fixed in advance. Repair opportunities are identical across conditions, and every attempted run contributes to cost.

Statistical estimand

Let $s_{qcr} \in \{0, 1\}$ be success for task q , condition c and repetition r . Define $\bar{s}_{qc}$ as its average over repetitions. The primary difference is

$$\Delta = \frac{1}{Q} \sum_{q=1}^Q (\bar{s}_{q,D3} - \bar{s}_{q,D1}). \quad (7)$$

The benchmark estimates average improvement over its explicitly sampled task inventory, not over all scientific research. Uncertainty must account for shared datasets and task templates. The planned analysis uses paired, stratified resampling of independent source/template families, retaining related tasks and all repetitions within each resampled family. A mixed-effects sensitivity analysis includes task family and dataset. If too few independent families are available, inference remains descriptive rather than treating near-duplicate tasks as independent.

The preregistration specifies family construction, resampling implementation, interval method, stopping rules and the minimum relevant improvement before evaluating the holdout. Development simulations determine whether the proposed inventory can resolve that improvement; the present article does not assert a powered sample size. Task-level results, denominators and uncertainty accompany any overall rate.

Complementary outcomes and failure gates

Secondary outcomes are numerical error, rerun agreement, unauthorized information use, unsupported-claim rate, citation support, expert correction time, latency and total computation cost. These are reported separately. A combined success score must not conceal that an analyst completes more tasks by producing more unsupported biological claims.

The D2–D1 and D3–D2 contrasts explain the role of semantics and lineage. Representation and model-choice comparisons are secondary unless assigned a separate registered hypothesis family. Multiple confirmatory contrasts require a declared familywise procedure; exploratory comparisons are labelled and carried forward to new data. A system cannot claim a reduction in rare harmful errors simply because a small benchmark observes none.

Information and resource matching

All agents use the same base model snapshot, tools, execution permissions, numerical library versions, maximum attempts and task-specific budget. Token exposure and consumed resources are reported even when equal maximum budgets produce unequal actual use. Differences in documentation necessarily alter the input; the length-matched control helps distinguish useful content from extra processing opportunity.

The researcher also evaluates a smaller or lower-effort model to test whether the contract’s benefit is model-dependent. A high-capacity language model is a candidate implementation, not part of the scientific definition. Exact provider and snapshot availability are recorded when the study runs. No superiority over alternative models is assumed from product positioning.

Negative controls for persuasive labels

Mechanism names can bias interpretation. In a prespecified audit, selected channels receive neutral identifiers while their numerical definitions remain available. A separate deliberately mismatched-label task tests whether the analyst follows the definition or a suggestive name. These are benchmark controls, never silent modifications to a real participant analysis.

Further controls remove an unconsumed upstream edge, swap history between matched participants, preserve temporal autocorrelation while breaking the tested alignment, and replace a proposed effect with a known null. Each perturbation has an explicit expected consequence. A graph edge deletion tests computation only if execution actually used the edge.

Conditions that would weaken the proposal

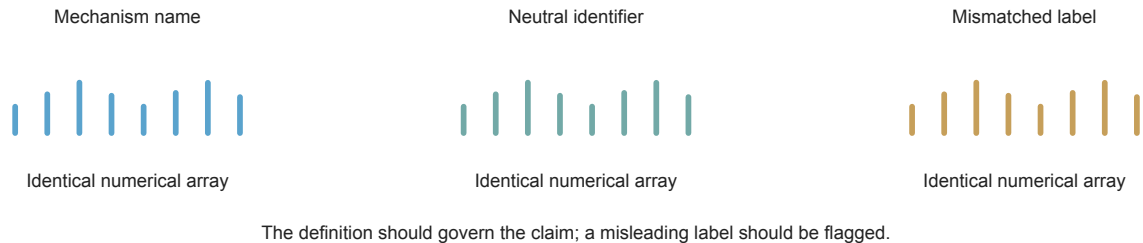
MILab’s central advantage is not supported if documentation increases confident stories but not valid completion, if gains vanish under equal resources, or if simpler acoustic descriptions perform equally well at lower cost. Failure to improve the primary contrast remains a publishable methodological result if the test is sufficiently informative. An advantage confined to one task family must be reported at that scope.

First application: learning musical pleasure

The companion listening design provides a possible application rather than the only purpose of MILab[18]. The current companion protocol uses exactly 20 recognized

Prespecified benchmark controls expose persuasive names and unconsumed graph edges.

a Same values and definitions, different labels

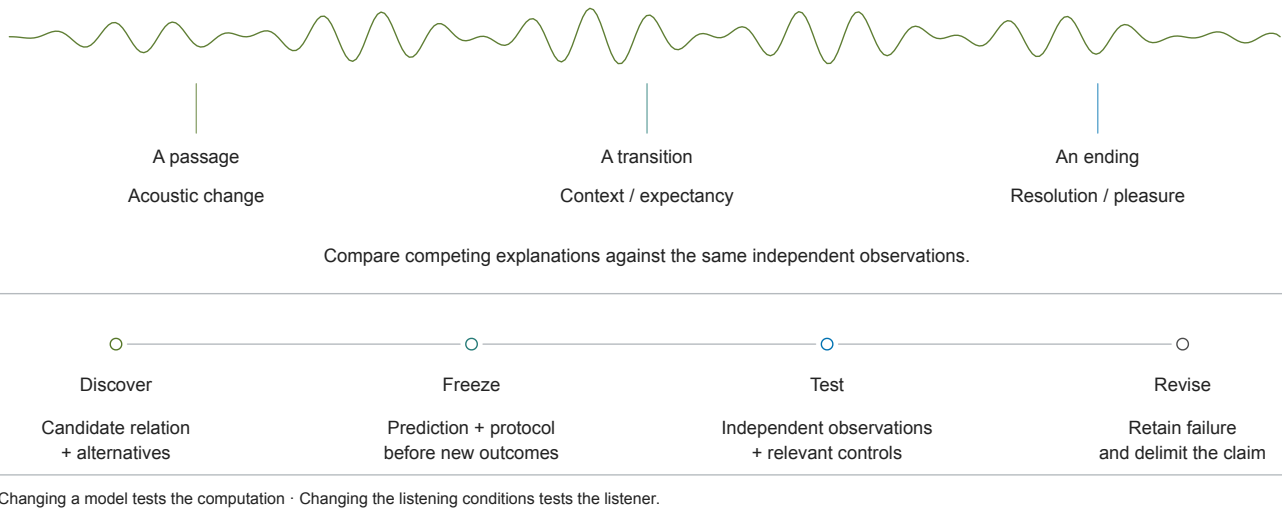


b A drawn edge matters only if the computation used it



These are labelled benchmark perturbations, never silent alterations of real participant analyses.

Fig. 10 | Persuasive labels are evaluated with explicit negative controls. Neutral and deliberately mismatched names accompany identical numerical arrays and available definitions in prespecified benchmark tasks. The analyst should follow operational definitions and identify misleading names. A separate dependency control removes an edge only under its declared audit; an unconsumed edge cannot establish a computational effect. These controls do not alter participant data, and no analyst success rate is shown.



Changing a model tests the computation · Changing the listening conditions tests the listener.

Fig. 11 | From a passage-level observation to a new empirical test. Distinct passages of one recording can supply competing acoustic and contextual descriptions. Discovery selects a candidate relation, not a confirmed explanation. A subsequent frozen prediction is tested with new observations or an appropriately controlled intervention. Changing MI's calculation and changing a listener's experience are different interventions.

classical composition families and explicit intake reports, followed by controlled and naturalistic tasks. It does not reconstruct general streaming history or autobiography. Other history sources would require a separate study specification. Recording identity and version must be verified before analysis.

The primary personal comparison asks whether history-conditioned predictions improve on a population mapping and whether earlier feedback adds further predictive value. The same responses also support carefully bounded comparisons of alternative acoustic, temporal and listening-related explanations. A failed direct pleasure channel does not automatically invalidate the representation; an exploratory substitute that fits the same data does not rescue it either. Its value must survive an independent test.

A worked question without invented results

Consider an analysis asking whether pleasure around an ending is better predicted by an acoustic-change account or by a context-dependent resolution account, and whether history changes that relationship. First, identify eligible endings independently of their pleasure outcomes and define the candidate computations. Second, specify the response interval and any motor lag using development information. Third, compare matched statistical models with common covariates and equally constrained tuning.

A resolution-related contribution is assessed relative to simpler acoustic predictors, repetition, familiarity and personal scale use. If candidate predictors remain too collinear to distinguish, the result is underidentification, not evidence that both corresponding brain mechanisms operate. A next experiment could vary preceding context while holding the ending fixed, or vary the ending within

a controlled context, provided the audible manipulation and remaining confounds are measured.

The next experiment is a separate commitment

The analyst may suggest stimuli where competing accounts disagree, including passages within the same work or controlled variants. A researcher reviews musical validity, rights, task burden and whether the manipulation changes more than the intended property. The new study commits predictions and analysis choices before its outcomes are available.

Adaptive stimulus selection is optional. If introduced, its policy, candidate set, randomization and selection probabilities must be logged, and the evaluation must address that selected sounds may not represent ordinary listening. The present baseline retains prepared sessions and offline analysis, avoiding an untested coupling between agent latency and participant experience.

MILab as a research system

Keep acquisition independent of the analyst

The operational laboratory separates catalogue preparation, participant sessions and analysis. Approved stimuli and schedules are released before a session. The browser presents audio and records observations with exposure identifiers and explicit timing. The analysis service consumes immutable releases. It cannot rewrite observations, change an already committed prediction or silently replace the protocol used to collect data.

The technical companion specifies this acquisition baseline, now named MI-Lab[19]. This article extends the analysis role beyond participant explanations or catalogue assistance. The LLM analyst operates on authorized research views and does not coach participants, invent consent decisions or provide missing responses. A prepared session can continue when the analyst service is unavailable.

Audio-context time, browser monotonic time and server receipt time remain separate. A delayed network upload does not alter when sound was presented. The accuracy of physical playback and response timing must be qualified on the actual devices, rather than inferred from timestamp resolution[20]. Replaying an excerpt creates another exposure; retrying an upload preserves the same event identity.

Data access is part of the analysis specification

Participant identity and contact details are kept outside the analyst’s research view. Pseudonyms do not make musical histories anonymous, and a summary table can still reveal a person in a small group. Each tool receives the minimum approved scope, and exports record their aggregation and suppression rules. A hosted LLM route must be documented in the consent and institutional data-flow review before participant information is transmitted.

MILab can run numerical workers within the institution and return approved summaries to the analyst. This reduces unnecessary raw-data transfer but does not guarantee anonymity. Sensitive interpretation should not become a participant-facing diagnosis. Autobiographical content, if collected, is a distinct data class with appropriate handling and a defensible measurement purpose.

Retrieved papers, participant text and imported meta-data are treated as evidence to inspect, not as instructions

that can change tool permissions. The execution environment uses bounded resources, isolated outputs and restricted network access appropriate to the task. Test outcomes remain outside the analyst’s accessible store until the evaluation service scores a committed artifact. Logging alone cannot compensate for an agent that was allowed to read its answers.

Retention, access and independent reconstruction

The public package should expose the scientific question, definitions, analysis specifications, relevant code, task scoring and permitted derived artifacts. Rights-restricted recordings and identifiable observations require controlled access or alternative reproducible fixtures. A data hash establishes identity but cannot enable reproduction without the underlying artifact.

MI’s existing architecture record states that some implementation records require controlled access[1]. MI-Lab must report which results can be reproduced publicly, by approved external auditors, or only internally. A claim of independent reconstruction requires an actual independent run. Publishing an interface or diagram does not meet that standard. FAIR stewardship permits access conditions when necessary, but those conditions and the route to reuse must be explicit[13].

Extending the lens to brain data

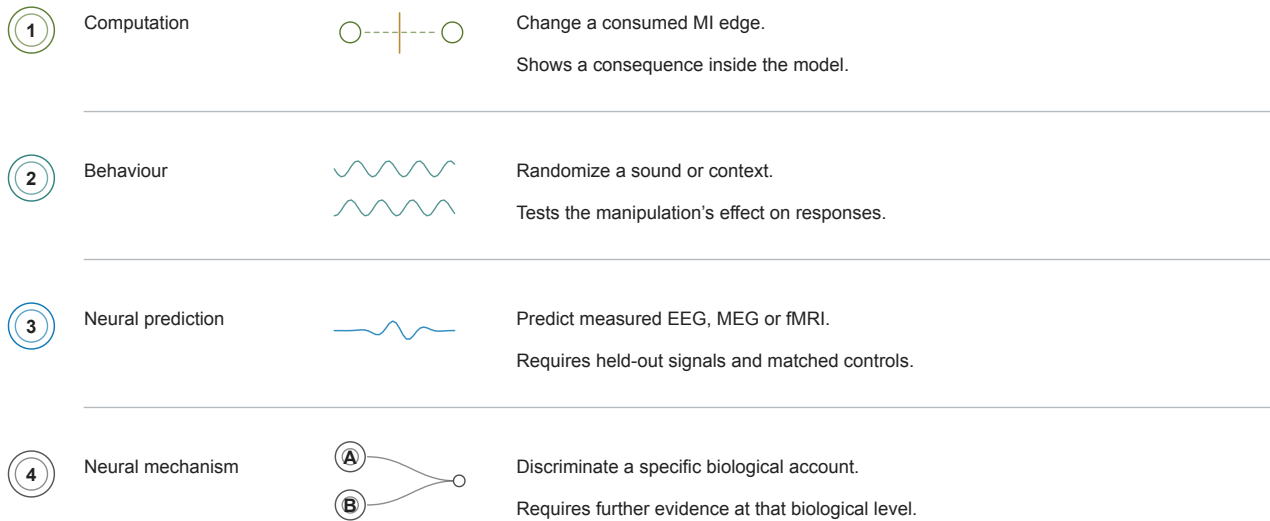
Behavioural pleasure is the first observation class, not the definition of the instrument. Neural or physiological measurements can be added through a separate observation model. For EEG or MEG, stimulus alignment, preprocessing, sensor/source definitions and temporal kernels must be recorded. For fMRI, haemodynamic modelling, run structure, motion handling and region selection become part of the analysis. Dataset conventions such as BIDS support consistent identity and metadata[21]; they do not establish a valid neural interpretation.

An encoding analysis asks whether a representation predicts measured responses under held-out evaluation. It does not establish that a named MI mechanism is implemented by the measured brain region. Simpler acoustic, timing and nuisance predictors must be evaluated under the same folds and tuning budget. Regions, time points and model variants selected for evaluation cannot be chosen using the same held-out responses used to claim success.

Recent work on language-model-to-brain comparisons illustrates the danger: shuffled splits and overlooked positional or rate signals can produce misleading alignment[22]. This finding is not a test of MI. It motivates analogous checks for autocorrelation, recording position, acoustic envelopes and analysis leakage in musical data.

The strongest extension would combine model discrimination with a controlled manipulation that makes competing accounts predict different outcomes. A change in an MI edge demonstrates a consequence within the computation. A change in behaviour following a randomized sound manipulation supports a claim about that manipulation under the design’s assumptions. A specific neural-mechanism claim requires further evidence matched to that level.

An inspectable equation and a good prediction are useful evidence at different levels.



No automatic promotion from a channel name to a brain mechanism.

Pleasure is the first observation class. Brain data need their own observation model and independent validation.

Fig. 12 | Computation, behaviour and neural explanation need distinct evidence. A consumed-edge perturbation tests the calculation. A randomized sound/context manipulation supports claims about that manipulation and measured behaviour under the design assumptions. Neural encoding requires independently measured signals, an observation model, held-out evaluation and matched alternatives. Establishing a specific biological mechanism requires further evidence at that level. The rows are distinct claim classes, not automatic stages of validation or measured results.

Discussion

The scientific role proposed here is broader than a single response predictor and narrower than an autonomous theory of the musical brain. MI supplies a revisable representation; MILab preserves observations and their provenance; the LLM assists analytical work. Their combination is useful if it helps researchers formulate distinctions, run sound comparisons and retain findings that survive independent tests.

This arrangement also changes what failure means. A direct readout can fail while a subset of the representation supports another valid account. Conversely, an accurate prediction can arise from low-level acoustics, stable rating habits or information leakage. The evidence contract makes these possibilities accessible; the comparison design determines whether they have been separated. Neither a rich architecture nor a powerful analyst removes the need for that design.

What remains interpretable

The framework preserves access to the selected MI computations and the numerical analysis that produced a result. It does not promise that the analyst's internal reasoning is transparent, that every feature has a unique psychological meaning, or that the original MI implementation is publicly reconstructable without its required records. These are different properties and should be reported independently.

Inspectability can still have substantial practical value. It permits a disputed normalization, temporal request or statistical assumption to be changed while retaining the rest of an analysis. It permits a reader to discover that two apparently independent mechanisms share most of their inputs. It permits a null result to be traced to an uninformative contrast rather than hidden by a persuasive ex-

planation. Whether these affordances improve research is the subject of the proposed benchmark.

Limits of the instrument metaphor

Unlike a microscope with a physical optical path, MI contains theory-guided transformations whose relation to cognition is itself under investigation. The metaphor is appropriate only if the instrument's transfer properties and limitations are characterized. A descriptor may be reliable as a computation yet invalid as a measure of memory or reward. A useful predictive representation may support several incompatible mechanistic explanations.

The repertoire can also shape discovery by making some questions easy and others invisible. Independent acoustic alternatives, neutral feature labels and new observation classes help reveal this selection. Cultural and musical breadth cannot be inferred from broad model names. A study based on familiar classical recordings supports conclusions about its actual sample and stimuli; additional styles, cultures, ages and listening contexts require corresponding evaluations.

Limitations of the proposed benchmark

The benchmark combines fixed-answer and open-ended tasks, which have different standards of correctness. Expert rubrics remain fallible, public replications may be contaminated by model training, and synthetic null cases cannot reproduce every empirical difficulty. Small numbers of independent datasets can make uncertainty large even with many runs. The benchmark's task inventory and all scoring choices must therefore be released or made auditable before its results are interpreted broadly.

Provider models can change, costs may constrain repeated runs, and archived prompts do not guarantee identical regenerated code. The primary reproducibility tar-

get is the frozen numerical analysis. Stability across fresh agent runs is measured separately. A high-effort model is a candidate research resource, not a substitute for validation or a permanent scientific dependency.

A feasible sequence of evidence

The first implementation milestone is one complete, independently rerunnable analysis on rights-compatible data, including a deliberate failure case. The next milestone is a blind comparison of numerical-only and evidence-contract conditions. A prospective listening study can then test the personal readouts and selected discoveries. Neural extensions follow when suitable measurements and expertise are available.

These stages provide increasingly demanding evidence without treating a software prototype as a participant study. They also allow the project to stop or change direction if the contract adds documentation but no measurable research benefit. A methodological paper can report that boundary as clearly as an improvement.

Conclusion

MI-Lab proposes an inspectable route from sound to computational descriptions, from descriptions to analyses, and from analyses to claims that can be challenged. Strong language models can help navigate that route while numerical execution, observational evidence and scientific responsibility remain explicit. The central empirical test is whether this arrangement improves the validity and reproducibility of research on music, cognition and the brain.

Declarations

Status and registration. This is a proposed methods framework and evaluation specification. It reports no new participant data, benchmark results, deployed hybrid service or completed preregistration. A final analysis plan, sample justification and institutional approvals are required before the corresponding studies.

Data, code and materials. The authoring package contains editable manuscript and figure sources, a source ledger and document-quality records. Figures are original schematics. It contains no participant dataset or functioning LLM-analysis service. Availability and access restrictions of the existing MI engine are described in its architecture record.

Authorship and interests. Amaç Erdem is the developer of MI and proposer of MI-Lab; this relationship should be considered when evaluating the framework. Final funding and competing-interest declarations require author confirmation before submission.

AI assistance. This draft was prepared with an OpenAI coding assistant for literature retrieval, manuscript drafting, figure programming and document checks. Those activities are not validation of the proposed analyst. The author must review and approve the scientific content before submission. No participant records were supplied during preparation.

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.

Appendix A. Minimum implementation contract

Proposed research objects and acceptance rules. These are specifications, not claims of an implemented database or completed validation.

Object	Minimum retained information	Acceptance rule
Stimulus release	Composition, performance, recording and excerpt IDs; waveform digest; decoding and trimming; rights record; channel layout and sampling rate.	A title match does not establish recording identity. Derived audio links to a permitted source and an explicit transformation.
MI quantity	Engine/configuration digest; channel identifier; equation or operational definition; units; normalization; consumed inputs; fallback status; temporal support.	Descriptor names do not supply missing units, probabilities or biological validity. Executed and declared dependencies remain distinct.
Observation	Pseudonymous participant, session, exposure and trial IDs; question wording and response scale; phase; timestamps; completeness and quality flags.	A repeated upload has one logical identity. Repeated listening creates a new exposure. Missing is distinct from zero.
Personal state	Intake version; eligible prior observations; learner specification; fitted state; seed; update time; state digest.	The current target and future reports are absent from the update. A history edit creates a new state version.
Analysis plan	Question; claim class; estimand; allowed features; exclusions; transformations; splits; tuning; multiplicity; uncertainty; failure policy.	Confirmation requires a pre-outcome commitment. Post-outcome revisions remain exploratory or receive a new independent test.
Execution	Frozen code; dependency/environment record; authorized data release; model configuration; tool calls; errors; runtime; output artifacts.	The analyst cannot change evaluation labels or acceptance tests. Rerun criteria state exact or tolerance-based numerical agreement.
Claim	Result pointer; estimate and uncertainty; population/stimulus scope; planned/exploratory status; source-role links; reviewer disposition.	Each empirical assertion is supported by its executed result. Neural interpretation cannot be inferred from a behavioural association alone.

Temporal eligibility

Eligibility is evaluated over the full transitive dependency graph, not just the final channel’s window. For a prediction at information time τ , every personal input must have been available before τ . Audio availability is specified separately: a whole recording can be an allowed offline stimulus input while still being inadmissible for a claim about causal online processing at a particular sound time.

An implementation should compute or conservatively bound each output’s earliest and latest input support, then

propagate those bounds through transformations. Unknown support is an explicit state. A model cannot gain eligibility by renaming an acausal descriptor or by storing it before a session.

Scientific acceptance

Numerical checks establish what the program computed. A scientifically valid conclusion also requires appropriate sampling units, comparison models, measurement definitions and interpretation. These checks are recorded separately so that disagreement can be located. A failed check

may invalidate a result, narrow a claim, or motivate another experiment; the response is not automatically to remove the offending observation.

Status transitions

A research object progresses through proposed, executable, internally checked, independently rerun and

empirically evaluated states only when the corresponding artifact exists. Deployment readiness and construct validity are separate statuses. Document-quality checks performed for this manuscript establish that the artifact can be read and inspected; they establish none of those research-performance states.

Appendix B. Domain-specific task cards

Illustrative task specifications. The tasks below are not a completed benchmark, and none has an observed pass rate.

Research question	Required distinction	What a valid answer must do
Does feedback help?	Fixed population, history-only and feedback-updated predictions use the same target and eligible prior information.	Score B–A and C–B separately; compare against a personal mean and equally informed acoustic learner; audit the prediction cutoff.
Is this channel surprise?	ICEM/PUPF-related audio scores, elicited probability and reported surprise have different definitions.	Retain the acoustic score’s definition; calculate information content only from a specified probability model; test correspondence rather than assume equivalence.
Is resolution informative?	A resolution-related descriptor can share inputs with roughness, energy or a generic change detector.	Test added held-out information against common acoustic controls; state when collinearity prevents discrimination; propose an independent contrast.
Is there a reward error?	A signed reported value discrepancy differs from sensory unexpectedness and from an RPEM-labelled audio output.	Keep these quantities separate; avoid validating an update by correlating algebraically coupled differences; predict later independent behaviour.
Is there memory?	Recorded exposure, recognition, recollection and a WMED-labelled calculation are different objects.	Restrict the conclusion to the measured construct; request a suitable memory assay for stronger claims; do not infer autobiographical content from a playlist.
Does MI predict the brain?	Neural encoding, selected-region association and a named biological mechanism make different claims.	Use genuine neural observations, appropriate nuisance baselines and held-out evaluation; prevent region/time selection on test responses.

Mechanism names preserve lineage

The task cards refer to names in the existing MI and listening-design records[1, 18]. ICEM denotes Information Content Emotion Model, PUPF Pleasure-Uncertainty Prediction Function, RPEM Reward Prediction Error in Music, and WMED Working Memory-Entrainment Dissociation. These names locate proposed computations; they do not certify that the output has the corresponding psychological or physiological meaning.

UDP, Uncertainty-Driven Pleasure, can be compared with PUPF as an alternative implemented account when their precise channels and support are specified. Their naming similarity is not a reason to collapse them, and different names are not a reason to assume independence. The analysis should expose shared acoustic and temporal inputs before assigning an incremental effect to either mechanism.

A useful failure report

If an analysis cannot distinguish two explanations, the output should identify the shared predictors, uncertain

measurements or missing contrasts responsible. If a discovery fails on held-out music, the report should retain that failure and describe its evaluation scope. If a computation cannot be reconstructed because records are unavailable, that is an access or provenance limitation; it is not a negative finding about the participant.

Release boundary

Before public evaluation, the task inventory, reference calculations, rubrics, family assignments, budgets, scoring code and development decisions must be frozen. Private evaluation labels remain inaccessible to analysts. After evaluation, release the allowable artifacts and failures together with an access statement for restricted data. The present authoring package establishes the manuscript and its design record, not this future benchmark release.