RefinedScience

Encoding Institutional Expertise: An Agentic Discovery Platform for Trustworthy Multimodal Analysis

Burke D[1], Riemondy K [1]

[1]RefinedScience, Aurora, CO

LEARN MORE
refinedscience.com

Institutional expertise is the missing input

Agentic workflows now reach every modality and orchestrate hundreds of tools, yet the scientists who produce a dataset cannot reliably make such systems follow the methodology behind it. That methodology is unwritten and dataset-specific: the unit of replication, which timepoints may be compared, which cohorts overlap, what a null result can claim. Expressed in a prompt it is advisory; it erodes as a session lengthens, disappears when the user or model changes, and leaves no record for review. Deprived of it, agents fall back on generic consensus answers that are confidently useless for discovery.

We present an AI co-scientist that treats institutional expertise as a first-class, engineered input. An ingestion pipeline captures the tacit methodology behind each dataset and compiles it into versioned, machine-enforceable constraints across three layers: the tools, the governing domain knowledge and the evaluation harness. Figures 1 to 4 describe the sources and compiled forms of that expertise, the surfaces at which it enters an analysis, a session as the user experiences it, and the classes of error it intercepts.

Expertise compiled into one enforceable layer

Stateless agents operate above this layer; the layer itself is persistent and versioned.

Language model

Expertise layer: institutional knowledge compiled into the tools, the governing domain knowledge and the evaluation harness

Analytical tools, external databases and institutional data

This layer is absent from systems that connect a model directly to data and tools.

***Institutional expertise** is the accumulated knowledge, human and encoded, that produced a dataset and governs how it is analyzed. It cannot be recovered from the data, only a fraction of it appears in the public record, and none of it accompanies the data when the data are provided to an agent.*

CONTRIBUTED BY	CATEGORY OF KNOWLEDGE	COMPILED INTO
• Computational biology (atlas construction) • Processing decisions embedded in the released data • Codebases, notebooks and SOPs	**Measurement provenance** How each modality was generated, normalized and denoised, and what that permits	Binding parameter defaults Typed interfaces
• Translational science • Biostatistics • Clinical hematology	**Study design** Which samples, timepoints and cohorts may validly be compared	Procedural preconditions Typed interfaces
• Clinical hematology and hematopathology • Clinical flow cytometry • Clinical data management	**Clinical semantics** What clinical variables, classifications and flow-cytometry gates mean	Controlled vocabularies Procedural preconditions
• Biostatistics • Computational biology (atlas analysis) • Codebases, notebooks and SOPs	**Statistical conventions** The unit of replication, multiplicity correction and power reporting	Binding parameter defaults Named checks
• Translational science • Academic collaborators and key opinion leaders • Peer-reviewed literature	**Evidentiary standards** What constitutes a finding, a replication and a validated result	Rubrics and ranking criteria Procedural preconditions
• Internal review records • Regression tests in the codebases • Peer-reviewed literature on known pitfalls	**Documented failure modes** Analytical errors previously encountered on these data	Named checks Independent oracles
• Translational science • Academic collaborators and key opinion leaders • Program leadership	**Scientific priorities** Which hypotheses merit investigation, and in what order	Validated procedures Rubrics and ranking criteria

Compiled forms are colored by the platform layer they populate: Tools; Governing domain knowledge; Evaluation harness

Figure 1. Sources and compiled forms of institutional expertise. Each category of knowledge is contributed by the sources listed at left and compiled into the enforceable forms at right; chip color indicates the platform layer each form populates.

An analysis is a trajectory; expertise enters at three surfaces

$$\tau = (q, a_1, o_1, \ldots, a_n, o_n, c) \quad a_i \sim \pi(\cdot \mid q, a_{<i}, o_{<i}) \quad a_i \in A \quad V(\tau, c) \in \{\text{accept, revise, reject}\}$$

q, question; *a*, action (a tool call); *o*, observation; *c*, claim together with its numbers and figures. An unconstrained agent relies on a prior π learned from public text; institutional judgment enters at the three surfaces below.

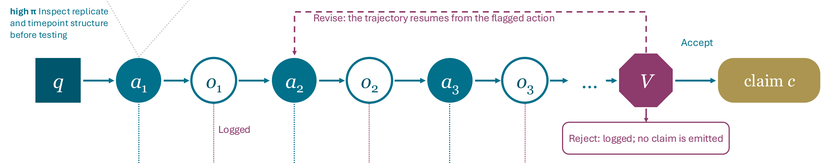
① POLICY PRIOR π · governing domain knowledge

Pool all cells across timepoints and test — low π

Rank targets by tumor expression alone — low π

high π Inspect replicate and timepoint structure before testing

Compiled domain knowledge conditions the policy prior on the dataset under analysis. The steps an expert would take next (inspecting replicate structure, matching timepoints, checking cohort overlap) receive high prior probability; branches known to be invalid receive low probability. Because the prior is implemented in the platform rather than in a prompt, it is identical for every user.



② ACTION SPACE A · tools and data

Cohort summary
Input: dataset manifest
Returns: fields, timepoints and replicate structure

Composition test
Unit of replication: donor (default)
Timepoints: matched
Multiplicity: corrected

Survival analysis
Cohorts deduplicated by patient
Minimum detectable hazard ratio reported

Tools are typed contracts. The unit of replication, timepoint matching, cohort deduplication and power reporting are parameters with enforced defaults, so an invalid analysis is difficult to express for any user or model. Data enter through the same mechanism, as registered manifests.

③ VERIFIER V(τ, c) · evaluation harness at run time

Provenance. Each number is hashed to the tool call and inputs that produced it (append-only log).

Validator. Identifiers and numbers (±1%) lacking a tool source are flagged.

Figure–text parity. Captions and reported numbers must agree with the rendered figure.

Reviewer. Inferences unsupported by the evidence are returned for revision.

Trajectory check. Tools, their order and their arguments are compared with the compiled procedure.

Surfaces ①–③ are implemented beneath the workflow rather than in its prompt; they therefore persist across users, models and sessions, and none of them is supplied by increasing model scale.

Figure 2. An analysis represented as a trajectory τ. Institutional expertise enters at three surfaces: a policy prior (governing domain knowledge), a typed action space (tools with contracts and statistical defaults) and a verifier over the trajectory and the claim. The path shown is schematic; no real query is reproduced.

Analyses are reproducible by construction

ANALYSIS — TRACE

USER

Which clinical variables are associated with a monocytic phenotype in the AML atlas? Test cell proportions against the mutation and risk panel.

1 The question is asked in plain language, and no code is written at any point in the session.

AI CO-SCIENTIST

Composition was tested per donor in diagnosis samples (n = 91; Welch's test on patient aggregates, BH-corrected). The strongest clinical associations fall on the differentiation axis around the monocytic state:

clinical axis	shift in composition	adj p
NPM1-mutant → GMP	+29.3 pp · 38.1% vs 8.8%	8e-4
FLT3-mutant → GMP	+14.1 pp · 24.8% vs 10.7%	0.012
ELN fav/int → GMP	+19.6 pp · 27.7% vs 8.1%	0.0015

The Monocytic state shows the canonical clinical immunophenotype on directly measured surface protein, with **CD14, CD64, CD11b and HLA-DR** markedly above every other lineage.

donor-level tests · turn 13 · table 02 · provenance ✓

2 Statistics are donor-level and timepoint-matched by default, identically for every user and model (Fig. 2).

3 Cell-state labels are anchored to controlled clinical vocabularies and verified against the flow immunophenotype.

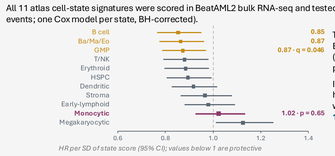
USER

Do these associations hold outside our cohort? Score the atlas cell states in BeatAML2 bulk RNA-seq and test overall survival.

4 Cohort definitions, filters and results carry forward across turns of the conversation.

AI CO-SCIENTIST

All 11 atlas cell-state signatures were scored in BeatAML2 bulk RNA-seq and tested against overall survival (591 patients, 355 events; one Cox model per state, BH-corrected).



The transferred signatures reproduce BeatAML2's published deconvolution (Monocytic vs Monocyte-like, **ρ = 0.93**; progenitor vs Progenitor-like, ρ = 0.82).

In the linked cohort (91 patients, 63 events), higher monocytic fraction is associated with worse overall survival **(HR 1.29 per SD, 95% CI 1.01–1.64, p = 0.039)**.

turn 29 · figure 05 · provenance ✓

5 Transferred signatures are validated against the external cohort's published deconvolution before interpretation.

6 A single session spans the single-cell atlas, an external bulk cohort and survival, with corrected statistics throughout.

7 Every number resolves to the traced call that produced it; figures and tables carry provenance tags.

Figure 3. A session as the user experiences it. Questions are asked in plain language and answered in conversation; composition is tested at the donor level, cell-state signatures are scored in an external bulk cohort and validated against its published deconvolution, and survival associations are reported with corrected statistics. The exchange is condensed from a recorded session, and all values are reproduced verbatim.

Each class of error is intercepted at a structural layer

A Intercepting layer by class of analytical error

	π domain knowledge	A tool contract	V validator	V reviewer	V trajectory	Information required
Pseudo-replication Cells rather than donors treated as independent observations	○	●			○	Dataset-specific
Timepoint confounding Diagnosis compared with remission; treatment effect read as biology	●	○				Dataset-specific
Untested read as negative Patients never assayed for a mutation counted as wild-type		●			○	Dataset-specific
Cohort double counting One cohort accessed through two portals and counted twice		●			○	Dataset-specific
Processing artifact as biology Denoised protein data in which every cell appears positive	●	○				Dataset-specific
Composition read as expression A whole-sample fold change that reflects cell-type composition	●			○		Domain judgment
Target called without a window Tumor expression reported without comparison to healthy tissue	●			○		Domain judgment
One signal counted three times Three prognostic associations that proxy a single latent axis				●	○	Domain judgment
Citation from memory A reference supplied without a retrieval step			●			Mechanical
Number without a source An identifier or value (±1%) not produced by any tool			●			Mechanical

● Intercepting layer ○ Secondary layer

Dataset-specific classes require facts that no public corpus contains: the donor identifier, the sampling design, the protein denoising procedure, the overlap between portals. Domain-judgment classes require an expert's procedure. Only the final two classes are mechanical, and increased model scale addresses only those.

B Nine independent layers of the evaluation harness

1	2	3	4	5	6	7	8	9
Tool unit tests	Ground-truth oracles (independent code path)	Validator: identifiers and numbers (±1%)	Trajectory checks: tools and arguments	Per-task rubrics on adversarial tasks	Scoring of multi-hypothesis runs	Efficiency budgets	Calibrated LLM judge	Golden-baseline regression testing

■ Grades the answer ■ Grades the reasoning ■ Grades efficiency

The harness grades the complete analytical trajectory against adversarial tasks, assessing both whether the conclusion is correct and whether expertise was applied at each step. No single check is trusted, and disagreement between layers is itself logged. On the current baseline, a 40-question golden set passes **97.8% of checks with zero fabrications**; six multi-agent runs comprising 832 tool calls produced **zero fabrications**; the calibrated judge scores **4.61 of 5 for accuracy** and **4.84 of 5 for calibration.**

Figure 4. (A) Each class of analytical error is intercepted by a structural layer, with a secondary layer behind it; the right-hand column indicates the kind of information required to detect the error. (B) The nine layers of the evaluation harness grade answers, reasoning and efficiency independently.

Conclusions

Expertise expressed in prompts cannot be enforced; expertise compiled into contracts, priors and checks can.

1 Institutional expertise is the accumulated knowledge that produced a dataset; compiled, it is the layer between a model and its data (Fig. 1).

2 The compiled rules hold across users, models, sessions and workflows (Fig. 2); to the user, a session is a question, traced tool calls and a report with provenance (Fig. 3).

3 A structural layer intercepts each class of analytical error, and nine independent checks grade reasoning as well as answers (Fig. 4).

FUNDING AND ACKNOWLEDGMENTS

RefinedScience is a joint venture between the University of Colorado Anschutz (CU Anschutz) and UCHealth. Data derive from CU Anschutz biobank specimens and linked UCHealth electronic medical record (EMR) data.

CONTACT

Devin Burke
devin.burke@refinedscience.com