Self-Revising Discovery Systems for Science: A Categorical Framework for Agentic Artificial Intelligence

Fiona Y. Wang
Laboratory for Atomistic and Molecular Mechanics
Department of Biological Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Markus J. Buehler
Laboratory for Atomistic and Molecular Mechanics
Department of Civil and Environmental Engineering
Department of Mechanical Engineering
Center for Computational Science and Engineering
Schwarzman College of Computing
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
mbuehler@mit.edu

Abstract

Scientific discovery is not only answer generation but revision of the representational regime in which evidence, artifacts, operations, and verifiers are typed. We develop a category-theoretic account of agentic discovery for materials science. In a fixed regime b with schema category $\mathcal{S}_b$, the system state is a copresheaf $I_t : \mathcal{S}_b \to \mathbf{Set}$, and provenance is the category of elements $\int_{\mathcal{S}_b} I_t$. Fixed-regime operation is an update on such states, endofunctorial only when provenance-preserving refinements are specified and preserved. Discovery is instead a verified regime transition $u : \mathcal{S}_b \to \mathcal{S}_{b'}$: old artifacts are preserved, transported by $\mathrm{Lan}_u I_t$, and compared with the post-transition state to identify residual content beyond functorial transport. This separates retrieval, search, and discovery without subjective novelty. We instantiate the framework in two systems. In Builder/Breaker, a protein-mechanics world model is revised under a Minimum Description Length gate; the accepted law expresses within-chain flexibility as all-mode elastic compliance conditioned by slow collective-mode participation, or mode-conditioned compliance. In CategoryScienceClaw, typed skills, artifacts, open needs, workflow mutation, gates, stress tests, and public discourse become a proof-carrying knowledge–computation graph. A fiber-network example records candidate models, rejected alternatives, an AIC gate, perturbation tests, and an accepted orientation-tensor anisotropic stiffness surrogate over an isotropic fiber-count descriptor. Together, the cases show how category theory can be both a mathematical language for discovery and an engineering specification for self-revising AI discovery systems.

Keywords: Agentic AI; scientific discovery; AI for science; category theory; regime transition; endofunctor; minimum description length; multi-agent systems; materials science; mechanics

1 Introduction

Artificial intelligence is now embedded in most stages of the scientific process. Foundation models retrieve and summarize literature, propose hypotheses, write and debug code, run and interpret simulations, design proteins and materials, and draft figures and reports. Agentic systems built on top of these models call external tools, coordinate multiple specialized subsystems, manage long-running workflows, and increasingly take partial responsibility for experimental decisions, both in computational pipelines and in autonomous laboratories [1–12]. The trajectory is clear and suggests that AI is no longer only a way to predict an output for a fixed task, but an active participant in how scientific work is structured.

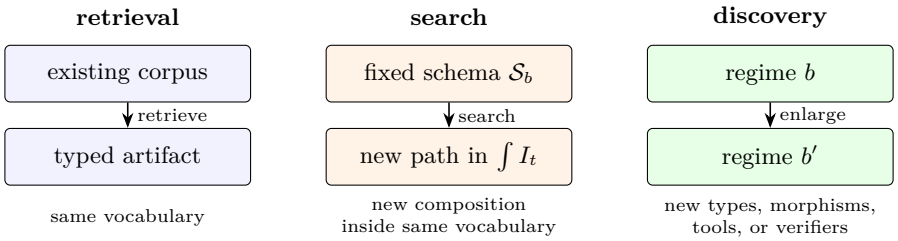


Figure 1: Retrieval, search, and discovery are structurally different operations. Retrieval adds an already representable artifact. Search finds a new path or object inside a fixed schema. Discovery changes the regime in which artifacts and operations are typed.

Yet the central question for these systems remains underformalized. Existing AI scientists are extraordinarily fluent at recombining, optimizing, and reformulating inside a fixed scientific vocabulary, but the operations that matter most in real science often change the vocabulary itself: a new effective variable, a new admissible operation, a new verifier, a new tool, a new artifact type. When is an agentic system searching within a fixed scientific regime, and when is it changing the regime itself? The answer is not merely philosophical; it determines how verifiers must be designed, how provenance must be audited, how progress should be measured, and why scaling a fixed model is qualitatively different from building a system that can construct new representational commitments.

This question has a lengthy and deep intellectual history. Popper emphasized critical tests and refutation; Kuhn emphasized changes of paradigm and world view; Lakatos described scientific progress through research programmes whose hard cores and auxiliary hypotheses evolve under pressure from anomalies [13–15]. The present paper extracts an operational problem from these earlier works: how can an artificial discovery system record, verify, and reuse the moment when evidence forces a change in the representational regime?

A concrete instance makes this distinction tangible. For this consider a researcher studying the mechanical response of a protein. They begin with a sequence, predict or retrieve a structure, construct a contact graph, diagonalize an elastic network, compare predicted fluctuations with crystallographic B-factors, formulate a hypothesis about which residues dominate the response, select another protein whose behavior should stress the hypothesis, and revise the model. The objects at every stage are typed: sequence, structure, contact graph, mode amplitude, feature, symbolic model, score, report. The operations are typed as well: build a contact graph from a structure, diagonalize a Kirchhoff matrix, extract a normal mode, fit a symbolic expression, score a description-length budget. The record of the work is not a string of answers but instead a typed provenance graph. Table 1 summarizes the corresponding implementation-to-categorical dictionary used throughout the paper.

Now suppose the next protein exposes a failure that cannot be repaired by changing a coefficient or adding another threshold. A local elastic-network feature may fit compact proteins but fail on hinge/domain proteins because the relevant phenomenon is no longer only local residue compliance; it is compliance expressed through a collective deformation. The researcher has two options. They can search within the current vocabulary, adjusting terms already available. Or they can enlarge the vocabulary by introducing a new effective type, operation, or verifier. The first move is search. The second is discovery in the strong sense used here: not only a better point in an existing space, but a change in the space of admissible scientific artifacts.

Category theory is useful here because it gives names to the engineering structure scientists already use. A schema is a category of artifact types and allowed operations. A current body of evidence is a population of artifacts over that schema. A provenance graph is the realized category of elements of that population. A consistent update is a natural transformation or functorial refinement. A discovery move is a transport from

Implementation object	Categorical object	Role in discovery system
artifact type	object of schema category $\mathcal{S}_b$	declares what kind of thing may be produced
tool or skill signature	morphism of $\mathcal{S}_b$	declares allowed transformation between types
current artifact population	copresheaf $I_t : \mathcal{S}_b \to \mathbf{Set}$	stores artifacts inhabiting each type
artifact DAG or hypergraph	category of elements $\int I_t$, or a multicategorical provenance analogue	realized provenance graph
accepted update	natural/refinement morphism	preserves prior typed provenance
gate or verifier	predicate or scoring functional	decides commitment, rejection, or supersession
new schema/tool/verifier	regime extension $u : \mathcal{S}_b \to \mathcal{S}_{b'}$	changes the admissible scientific vocabulary

Table 1: Dictionary between implementation terms and the categorical formalism used in this paper.

one schema to a larger or different schema, preserving what was valid while making new types, morphisms, tools, or verifiers available. The categorical foundations are standard; the applied use of categories for scientific schemas, ologs, and data migration is already well developed [16–22]. The language is abstract, but the objects are practical: PDB chains, simulations, equations, hypotheses, artifacts, claims, and reports.

This paper also continues a materials-science arc developed in our earlier work on ologs, hierarchical materials, learned maps, neural ologs, language-mediated reasoning, and plannerless scientific swarms [23–27]. In each case the scientific problem is not only to compute an output, but to preserve the structure that makes the output meaningful across scales. The present work turns that arc back onto discovery itself. If a material is a hierarchy of composable mechanisms, an agentic discovery system is a hierarchy of composable scientific artifacts.

We note that the underlying claim is not ordinary reductionism. In hierarchical materials, complexity is rarely located at a single privileged scale. Hydroxyapatite chemistry, collagen structure, mineral organization, crack-tip mechanics, and tissue remodeling are each necessary, but none is sufficient alone. What matters is the compositional grammar that organizes simple components into higher-order structure, and the responsiveness by which that structure updates under load, damage, growth, or new evidence. This view has predecessors in older scientific accounts of form and transformation, including Goethe's morphology [28]. Category theory is useful because it describes the morphisms among scales without forcing the explanation to live only at the bottom. The same principle motivates the present account of discovery: a scientific AI system should not only optimize artifacts inside a fixed representation, but should compose typed artifacts across representational levels, test those compositions against the world, and revise the grammar when the old one is too small.

This lineage also reverses the usual direction of influence between artificial intelligence and mechanics. In many current settings, AI is introduced into mechanics as an external optimizer or surrogate predictor: it accelerates simulation, fits constitutive laws, or searches a design space. The present framework instead lets mechanics help define what an AI discovery system should be. Mechanics supplies a disciplined language for state, load, response, instability, failure, admissible motion, constitutive closure, and multiscale transfer. These ideas reappear here as artifact states, evidence pressure, stress tests, gates, regime transitions, residual content, and provenance-preserving transport. In this sense, mechanics is not only a domain on which the framework is demonstrated but instead is one of the sources of the framework. The same habits that make a mechanical model meaningful (e.g., tracking boundary conditions, preserving invariants, testing failure modes, coarse-graining across scales, and distinguishing a new constitutive structure from a refit of old variables [29–36]) become design principles for agentic AI systems that revise their own scientific vocabulary.

We make four contributions. First, we give a formal semantics for typed artifact states as copresheaves $I_t : \mathcal{S}_b \to \mathbf{Set}$ and for realized provenance as the category of elements $\int_{\mathcal{S}_b} I_t$. This also gives a categorical generalization of the scientific knowledge graph: knowledge, computation, verification, rejection, public discourse, and schema revision are represented as parts of one executable, self-revising knowledge–computation graph rather than as separate data, workflow, and communication layers. Second, we distinguish fixed-regime agentic updates, modeled as endofunctorial dynamics under explicit assumptions, from discovery moves, modeled as verified regime transitions equipped with Kan-extension transport and an explicit preservation

map for old evidence. The empty value of the Kan extension on isolated new types gives a concrete obstruction: transport alone cannot populate them. Third, we use the Builder/Breaker protein-mechanics system as a quantitative case study in which a symbolic world model is revised under a Minimum Description Length (MDL) gate [10]; the accepted law expresses within-chain protein flexibility as all-mode elastic compliance conditioned by slow collective-mode participation, a "mode-conditioned compliance" relation that appears as a newly admitted interaction type rather than an additional term. Fourth, we use CategoryScienceClaw as a categorical layer on ScienceClaw in which the skill registry, immutable lineage, pressure coordination, workflow mutation, and public discourse are lifted into typed objects, morphisms, open needs, proof certificates, and audit records. The worked example in this paper is a mechanics investigation whose accepted models, rejected alternatives, gates, stress tests, and regime-transition claims are materialized as typed artifacts and morphisms, then rendered into human-readable scientific figures. In that example the committed mechanics object is an orientation-tensor anisotropic stiffness surrogate, accepted over an isotropic fiber-count descriptor by an AIC gate, so that model selection itself, including the rejected alternative, is recorded as typed provenance. Formal definitions are gathered in the Materials and Methods section; the Results section develops their practical interpretation through these two case studies.

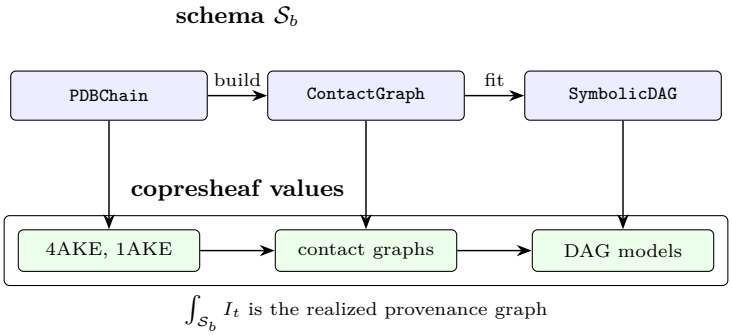


Figure 2: A fixed regime has a schema category $\mathcal{S}_b$ of types and operations. A copresheaf $I_t : \mathcal{S}_b \to \mathbf{Set}$ assigns actual artifacts to each type. The category of elements $\int I_t$ is the realized typed artifact DAG.

2 Results and Discussion

2.1 Agentic discovery systems are typed artifact systems

An agentic discovery system is best understood as a typed artifact system. Its persistent state is not a conversation transcript, a hidden vector, or a single model checkpoint. It is a growing record of artifacts and their provenance: data, simulations, models, hypotheses, code, measurements, reports, critiques, and decisions. Each artifact has a type, and each operation has a declared source and target type. Critically this typing is not bureaucratic overhead but what distinguishes a scientific claim from a fluent answer, addressing an important limitation of many probabilistic AI systems.

Five components recur across the systems considered here.

1. **A schema of artifact types and operations.** Types include sequences, structures, contact graphs, trajectories, hypotheses, symbolic models, measurements, and reports. Operations include structure prediction, simulation, normal-mode extraction, retrieval, proof checking, dimensional analysis, fabrication, and scoring.
2. **A population of artifacts over that schema.** The system stores actual sequences, structures, simulations, equations, figures, claims, and reports inhabiting the declared types.
3. **A provenance graph.** Every accepted artifact records its parents and the operation that produced it. Composition of operations is the scientific lineage from raw input to claim.
4. **A gate or verifier.** New artifacts are not automatically committed. They are accepted, rejected, superseded, or held for review by an explicit gate. MDL is one such gate; pressure scoring, schema overlap, peer review, and community feedback are others.
5. **A regime-update mechanism.** When evidence cannot be represented in the current schema, the system must extend or revise the schema, grammar, verifier, or tool registry.

The distinction among retrieval, search, and discovery is shown schematically in Fig. 1. Retrieval adds artifacts already expressible in the schema. Search explores combinations of existing artifacts and operations. Discovery changes the regime by adding or revising the types and operations under which future artifacts are judged.

This view accommodates both compact experimental systems and large distributed systems. In ProtAgents, the schema contains sequences, structures, force predictions, and protein-design hypotheses [3]. In Sparks and SciAgents, it contains research goals, generated hypotheses, executable plans, code outputs, and reports [4, 5]. ScienceClaw × Infinite instantiates the same mathematics at a larger architectural scale [9][1]. ScienceClaw supplies the execution substrate: an extensible registry of typed scientific skills, immutable artifacts with metadata and parent lineage, shared open needs, plannerless coordination, pressure scoring, and mutation of the active artifact graph. Infinite supplies the discourse substrate: structured posts, hypotheses, methods, findings, links among claims, votes, comments, reputation, and moderation. Together they make scientific work auditable across both computation and communication. In CategoryScienceClaw, that ScienceClaw substrate is lifted into typed categorical state: the schema contains typed skills, immutable artifacts with parent lineage, pressure-ranked open needs, workflow mutation records, public discourse objects, domain inputs, descriptors, candidate models, accepted and rejected alternatives, gates, stress tests, regime-transition records, figures, and reports. The objects differ, but the structural skeleton is the same.

2.2 Artifact states are copresheaves

The clean mathematical object behind the typed artifact graph is a copresheaf (Definition 2). Fix a discovery regime b (Definition 1). The regime includes a schema category $\mathcal{S}_b$: its objects are artifact types and its morphisms are allowed operations. A scientific state at time t is a covariant Set-valued functor

$$I_t : \mathcal{S}_b \longrightarrow \mathbf{Set}. \tag{1}$$

For each type A, the set $I_t(A)$ contains the artifacts of type A currently available to the system. For each operation $f : A \to B$, the function $I_t(f) : I_t(A) \to I_t(B)$ records how an artifact of type A gives rise to an artifact of type B, when that operation has been realized.

For a materials scientist, typical fibers are:

$$I_t(\text{PDBChain}), \qquad I_t(\text{ContactGraph}), \qquad I_t(\text{SymbolicDAG}),$$

the current PDB chains, the contact graphs constructed from them, and the candidate or accepted symbolic models. The actual provenance DAG is recovered by the category of elements $\int_{\mathcal{S}_b} I_t$ (Fig. 2). Its objects are pairs (A, x), where A is a type and $x \in I_t(A)$ is an artifact of that type. A morphism $(A, x) \to (B, y)$ is an operation $f : A \to B$ such that $I_t(f)(x) = y$. Thus the category of elements is not a metaphor for provenance. It is the typed provenance graph.

$$\begin{array}{ccccc} I_t & & \xrightarrow{\quad\delta_t\quad} & & I_{t+1} \\ {\scriptstyle \mathrm{Lan}_u}\downarrow & & & & \downarrow{\scriptstyle \mathrm{Lan}_u} \\ \mathrm{Lan}_u\, I_t & & \xrightarrow[\mathrm{Lan}_u(\delta_t)]{} & \mathrm{Lan}_u\, I_{t+1} & \xrightarrow{\bar{\rho}_{t+1}} I'_{t+1} \end{array}$$

Figure 3: Fixed-regime operation is the update Φ_b inside a schema $\mathcal{S}_b$. A committed fixed-regime step is represented by a refinement $\delta_t : I_t \to I_{t+1}$ associated with the object-level update $I_{t+1} = \Phi_b(I_t)$. The lower dashed arrow is $\mathrm{Lan}_u(\delta_t) : \mathrm{Lan}_u I_t \to \mathrm{Lan}_u I_{t+1}$, not an independent new-regime dynamics. Thus the left square commutes by functoriality of Lan_u on refinement morphisms. Discovery enters through the comparison map $\bar{\rho}_{t+1}$ from transported evidence to the verified post-transition state (Definition 6). In general $\mathrm{Lan}_u\, I_{t+1} \neq I'_{t+1}$; the comparison map and the residual artifacts outside its image record what the discovery move added beyond functorial transport.

Copresheaves are the right level of abstraction because they separate a regime from its current contents. The schema can remain fixed while the artifact population grows. Conversely, a discovery move can change

[1] Code at `https://github.com/lamm-mit/scienceclaw`, `https://github.com/lamm-mit/infinite`, and the CategoryScienceClaw mechanics branch at `https://github.com/lamm-mit/scienceclaw/tree/categoryscienceclaw-mechanics`.

the schema while preserving the old artifact population by transport into a new regime. This separation is exactly what is missing when all artifacts, operations, and updates are collapsed into one untyped graph.

This construction can be understood as a typed generalization of a scientific knowledge graph. A conventional knowledge graph records entities and relations over a largely fixed ontology. Here the fibers $I_t(A)$ contain scientific artifacts; morphisms $f : A \to B$ are executable or audited operations; the category of elements records realized provenance; gates record commitment and rejection; and regime transitions allow the vocabulary itself to change.

For a formal definition, fix a discovery regime

$$b = (\mathcal{S}_b, \Gamma_b, V_b, L_b).$$

A categorical knowledge–computation graph at time t is the tuple

$$\mathfrak{K}_t^b = \big(\mathcal{S}_b, \Gamma_b, I_t, \mathsf{Prov}_t, V_b, L_b, \mathsf{D}_t, \pi_t\big),$$

where $I_t : \mathcal{S}_b \to \mathbf{Set}$ is the typed artifact state; Prov_t is the realized provenance object, equal to $\int_{\mathcal{S}_b} I_t$ in the unary case and to the corresponding typed multicategory or colored-operadic provenance object when artifacts have multiple parents; V_b is the verifier or gate; L_b is the description-length or model-selection functional when present; D_t is an optional discourse category of claims, posts, evidence bundles, objections, replications, and validation signals; and π_t is the publication map carrying audited artifact paths or hyperpaths into public claim objects when such a discourse layer is implemented.

A conventional scientific knowledge graph is recovered by fixing $\mathcal{S}_b$, forgetting $\Gamma_b, V_b, L_b, \mathsf{D}_t, \pi_t$, and viewing Prov_t only as an entity-relation graph. A workflow-provenance graph is recovered by retaining production edges but not the full gate, rejection, discourse, and regime-transition structure. Hence $\mathfrak{K}_t^b$ is not simply a knowledge graph with metadata but instead an executable, verifier-aware, provenance-preserving scientific state. Fixed-regime search is evolution of $\mathfrak{K}_t^b$ under Φ_b; discovery is a verified transition

$$\mathfrak{K}_t^b \longrightarrow \mathfrak{K}_{t+1}^{b'}$$

with transported evidence $\mathrm{Lan}_u I_t$ and residual content outside $\mathrm{im}(\bar{\rho})$.

Remark 1 (Yoneda reading)**.** For a type A, the representable copresheaf $\mathcal{S}_b(A, -)$ encodes all operations that can be applied to a generic artifact of type A. An actual artifact $x \in I_t(A)$ induces, by the covariant Yoneda lemma, a natural transformation $\mathcal{S}_b(A, -) \to I_t$: each operation out of A is sent to the artifact obtained by applying that operation to x. In implementation terms, an artifact is known by the typed operations it can participate in and the downstream artifacts those operations produce.

2.3 Fixed-regime updates are endofunctorial under explicit assumptions

Inside a fixed regime b, an agentic system updates artifact states. Write $[\mathcal{S}_b, \mathbf{Set}]$ for the functor category of copresheaves on $\mathcal{S}_b$. Abstractly, a fixed-regime update is an operation on copresheaves,

$$\Phi_b : [\mathcal{S}_b, \mathbf{Set}] \longrightarrow [\mathcal{S}_b, \mathbf{Set}]. \tag{2}$$

It reads the current artifact population, selects compatible operations, proposes new artifacts, applies a gate, and returns the next artifact population. The statement that Φ_b is an endofunctor is a further preservation claim: after the category of admissible refinements between artifact states is specified, Φ_b must extend from an object-level update to a map on refinement morphisms in the corresponding category of knowledge states $\mathcal{K}_b$ (Definition 4). For the structural observations and the Kan proposition, $\mathcal{K}_b$ is the subcategory of $[\mathcal{S}_b, \mathbf{Set}]$ in which morphisms are componentwise injective natural transformations (Methods, Section 4); refinements may add, annotate, or supersede artifacts, but may not silently identify two previously distinct accepted artifacts.

This qualification matters. A raw program that takes a JSON artifact ledger to another JSON artifact ledger is only an endomap. It becomes an endofunctor only when it also preserves refinement morphisms: if one artifact state extends another by adding verified artifacts without overwriting prior provenance, the updated state should extend the updated predecessor in the same way. This is the formal version of a familiar engineering requirement: if a pipeline is refactored, old valid workflows must still compose.

In code, this is an audit contract rather than a slogan. The implementation must maintain stable artifact identifiers, typed tool or skill signatures, explicit parent lineage, append-only or explicit supersession semantics, status records for failed or retried calls, and no silent merge or deletion of accepted artifacts. When these conditions hold at the committed-state layer, a refinement $\alpha : I \to J$ can be pushed through the update to give a refinement $\Phi_b(\alpha) : \Phi_b(I) \to \Phi_b(J)$. When they fail, the system may still be useful, but the endofunctor model applies only after adding the missing audit structure.

Real systems may be stochastic because agents sample, tools fail, schedulers branch, and human feedback arrives asynchronously. Then Φ_b should be read as a relation, a stochastic kernel, or a morphism in a Kleisli category for an appropriate probability monad [37]. This view accommodates sampling-based agents, noisy verifiers, and partial reductions of the artifact graph without changing the structural claims of the framework. The deterministic notation is used because it displays the structural claim clearly; the categorical account is not tied to determinism.

This distinction is important in the examples below. Builder/Breaker is closest to the strict model at the layer of MDL-accepted symbolic DAGs and accumulated evidence, not at the raw proposal trace. CategoryScienceClaw is close at the mechanics-audit layer because typed skills, immutable artifacts, parent lineage, candidate models, accepted and rejected alternatives, gates, stress tests, regime transitions, reports, and public discourse are recorded explicitly; a fully endofunctorial implementation would further require checked schema signatures and explicit refinement maps between artifact states.

2.4 Discovery is regime transition, not only iteration

Search is iteration of Φ_b inside a fixed regime. Discovery requires a transition to a new regime (Definition 6). Let

$$u : \mathcal{S}_b \longrightarrow \mathcal{S}_{b'} \tag{3}$$

be a schema map from the old regime to the new one. In the simplest case, u is an inclusion that preserves old types and operations while adding new ones. In more realistic cases, it may also refine old types, split a type into subtypes, add a verifier, or add new morphisms between old objects. Therefore the general notion should not be restricted to a fully faithful embedding of categories; discovery may add new admissible relations among old objects.

The transition is summarized in Fig. 3. The old artifact state is transported into the new schema by left Kan extension:

$$\mathrm{Lan}_u I_t : \mathcal{S}_{b'} \longrightarrow \mathbf{Set}. \tag{4}$$

For an object A' of the new schema, the value $(\mathrm{Lan}_u I_t)(A')$ is computed as a colimit over the old artifacts that map toward A' (the precise formula appears as Eq. 9 in Methods). Operationally, it is the least systematic way to reinterpret old artifacts inside the new vocabulary. If A' receives no morphisms from the image of u, the comma category indexing the colimit is empty, so $(\mathrm{Lan}_u I_t)(A') = \varnothing$: free transport supplies nothing at that isolated new type. If A' does receive a morphism from an old type, then old evidence can be transported to it, even if A' itself is a new object. A discovery move is therefore not merely $\mathrm{Lan}_u I_t$; it is transport plus a verified post-transition state containing new evidence, new artifacts, new verifier outcomes, or new grammar productions that are not accounted for by transport alone.

This gives a precise form to the slogan that discovery changes the world model. Yet importantly it does not mean the old world disappears. The old artifacts persist as transported evidence, and their old provenance must remain auditable. What changes is the regime in which the evidence can be represented and composed.

The Kan-extension picture also gives a quantitative reading of how much discovery has occurred, but only after one additional datum is specified. A verified transition includes a natural transformation

$$\rho : I_t \longrightarrow u^* I'_{t+1},$$

where u^* restricts a new-regime state back to the old schema. This map says, explicitly, how each old accepted artifact is preserved in the new state. By the adjunction $\mathrm{Lan}_u \dashv u^*$, ρ corresponds uniquely to a comparison map

$$\bar{\rho} : \mathrm{Lan}_u I_t \longrightarrow I'_{t+1}.$$

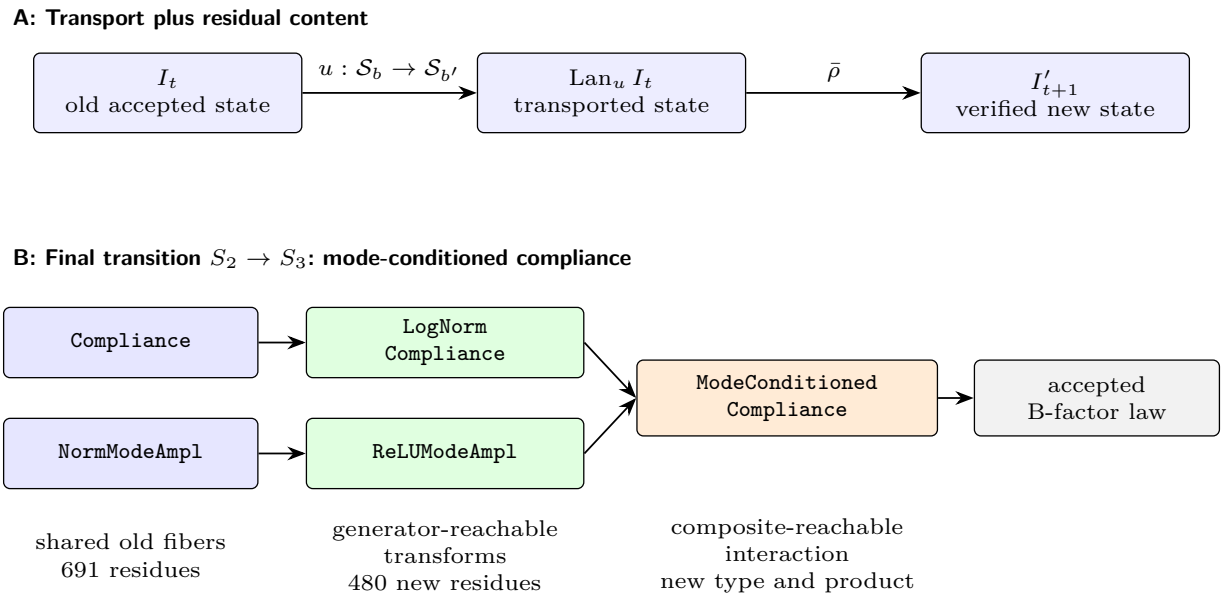


Figure 4: Kan-transport audit of the Builder/Breaker protein-mechanics run. (A) A verified transition transports the old artifact state by Lan_u and compares it with the accepted new state by $\bar{\rho}$; residual content records what is added beyond functorial transport. (B) In the final transition, log-compliance and shifted ReLU mode participation are generator-reachable transformations of old physics-derived quantities, whereas `ModeConditionedCompliance` is reachable only after the new regime admits a product operation. This is the categorical form of the mechanics insight in Eq. 7.

The image of $\bar{\rho}$ is the transported-evidence substate. Artifacts outside this image, object by object, are the empirical or representational content the system had to acquire beyond functorial reinterpretation of old evidence. When the new regime carries a relative description-length functional $L_{b'}(- \mid -)$, the **discovery cost** is the bit budget $L_{b'}(I'_{t+1} \mid \mathrm{im}(\bar{\rho}))$ required to specify the post-transition state given transported evidence. The Builder/Breaker model in Section 2.5 uses the MDL gate to measure this kind of cost in a concrete protein-mechanics regime.

2.5 The Builder/Breaker model gives a quantitative MDL case

The Builder/Breaker protein-mechanics model is a compact empirical instance of the framework, building on preliminary work reported in [10] (Definition 5[2]). The Breaker chooses new proteins intended to expose failure modes of the current symbolic model. The Builder proposes symbolic DAG edits. The gate accepts a candidate only if it reduces total description length after paired refitting on the same accumulated evidence:

$$L(M, D) = L_{\mathrm{model}}(M) + L_{\mathrm{data}}(D \mid M). \tag{5}$$

Here M is the symbolic DAG world model and D is the accumulated protein-mechanics evidence. If the Breaker adds stress-test evidence E, a proposed revision M' becomes part of the world model only when

$$L(M', D \cup E) < L(M, D \cup E),$$

after both models have been judged on the same evidence. This is the formal sense in which a productive failure becomes scientific structure: the new symbolic law must explain the counterexamples well enough to pay for its additional bits. Consequently, the acceptance criterion in this case is not monotone improvement of a single predictive score across iterations. Each outer iteration changes the accumulated evidence set on which the current and proposed symbolic DAGs are compared. The relevant paired test is therefore whether the revised model compresses the enlarged evidence set better than the previous model after both are refit on that same evidence, not whether the reported R^2 forms a monotone sequence across heterogeneous stages.

The relevant fixed schema contains PDB chains, Cα coordinates, contact graphs, Kirchhoff matrices, GNM spectra, compliance observables, slow-mode observables, feature terms, symbolic DAGs, B-factor targets,

[2]Code at `https://github.com/lamm-mit/BreakingTheWorld`.

and MDL budgets. The physical base is the elastic-network and Gaussian Network Model tradition, where low-resolution harmonic contact networks explain slow collective motions and residue-level temperature factors from structure alone [38–40]. For a protein chain p with residues $i = 1, \ldots, N_p$ and Cα coordinates $\mathbf{r}_{pi} \in \mathbb{R}^3$, define the contact graph by

$$A_{ij}^{(p)} = \mathbf{1}\{i \neq j,\ \|\mathbf{r}_{pi} - \mathbf{r}_{pj}\| < r_c\}, \qquad r_c = 10.0 \text{ Å}.$$

The GNM Kirchhoff matrix is

$$\Gamma_{ij}^{(p)} = \begin{cases} -A_{ij}^{(p)}, & i \neq j, \\ \sum_{k \neq i} A_{ik}^{(p)}, & i = j. \end{cases}$$

Diagonalize

$$\Gamma_p \mathbf{u}_{pk} = \lambda_{pk} \mathbf{u}_{pk}, \qquad 0 = \lambda_{p1} \leq \lambda_{p2} \leq \cdots \leq \lambda_{pN_p}.$$

The all-mode compliance of residue i is the diagonal of the pseudoinverse,

$$C_{pi} = (\Gamma_p^+)_{ii} = \sum_{\lambda_{pk} > 0} \frac{u_{pik}^2}{\lambda_{pk}}.$$

For connected contact graphs this is the familiar sum over $k = 2, \ldots, N_p$; the positive-eigenvalue form is the correct pseudoinverse expression if a cutoff or chain segmentation produces additional zero modes. This is not an arbitrary learned feature but instead the harmonic-network compliance implied by the residue contact topology. In the standard GNM relation,

$$\langle \Delta R_{pi}^2 \rangle = \frac{3 k_B T}{\gamma} C_{pi}, \qquad B_{pi}^{\mathrm{GNM}} = \frac{8\pi^2}{3} \langle \Delta R_{pi}^2 \rangle,$$

where γ is the effective spring constant. Because the learning target is normalized within each chain, the global constants T, γ, and $8\pi^2/3$ drop out. The experimental target is the per-chain normalized Cα crystallographic B-factor,

$$B_{pi}^{(z)} = \frac{B_{pi} - \overline{B}_p}{s_{B,p}},$$

so the model explains within-chain flexibility patterns rather than absolute crystallographic scale. The physically important caveat is that crystallographic B-factors include thermal motion, static disorder, refinement effects, and crystal-packing effects; the discovery task is therefore to find a compact structural proxy for the experimental pattern, not a complete molecular dynamics law.

Let $z_p(x_i) = (x_i - \overline{x}_p)/s_{x,p}$ denote per-chain normalization. The log-compliance feature and slowest collective-mode participation are

$$\phi_{pi} = z_p(\log(C_{pi} + \epsilon)), \qquad \psi_{pi} = [z_p(|u_{pi2}|) + \theta]_+ ,$$

where $[x]_+ = \max(x, 0)$, $\epsilon > 0$ is a small numerical floor inside the logarithm, and u_{p2} denotes the first nonzero GNM eigenmode, called `mode1_abs_z` in the run after taking absolute value and z-scoring. If a contact graph has more than one zero mode, u_{p2} should be read as the first positive-eigenvalue mode. The shift θ is added before the ReLU, so it acts as a lower clip rather than a conventional threshold; the equivalent threshold form is $[z_p(|u_{pi2}|) - \tau]_+$ with $\tau = -\theta$, and the fitted $\theta = 2.2678$ corresponds to clipping near the lowest observed value of $z_p(|u_{pi2}|)$ on the mixed validation slice, so ψ_{pi} shifts first-mode participation onto a positive scale and zeroes only the residues that barely participate in the dominant collective deformation. The discovered symbolic law has the form

$$\boxed{\widehat{B}_{pi}^{(z)} = \alpha + \beta\, \phi_{pi} \psi_{pi}} \tag{6}$$

or, expanded in the GNM spectrum,

$$\boxed{\widehat{B}_{pi}^{(z)} = \alpha + \beta\, z_p\left(\log\left(\sum_{\lambda_{pk}>0} \frac{u_{pik}^2}{\lambda_{pk}} + \epsilon\right)\right) \times \left[z_p(|u_{pi2}|) + \theta\right]_+ .} \tag{7}$$

The fitted numerical values accepted by the MDL gate are $\alpha = -0.1332$, $\beta = 0.2239$, and $\theta = 2.2678$. The important discovery event is not merely the use of a normal mode; normal modes are available from the start. Categorically, this is not the appearance of an isolated new object. It is the transition at which the schema admits a new multi-input morphism

$$\texttt{LogNormCompliance} \times \texttt{ReLUModeAmpl} \longrightarrow \texttt{ModeConditionedCompliance},$$

whose target is composite-reachable from old physics-derived quantities only after the new unary transforms and product operation are admitted. The audit below makes this composite reachability explicit. Mechanically, the resulting interaction supports a new explanatory role: local softness matters most when it is expressed along a global collective deformation, rather than acting as an independent additive cause.

For additional context the four outer iterations are summarized in Fig. 5. Iteration 0 fits a minimal local fluctuation model on compact proteins. Iteration 1 adds boundary and slow-mode structure for terminal flexibility and gains 9.0 bits. Iteration 2 exposes a hinge/domain-motion regime, using open and closed adenylate kinase (PDB chains 4AKE and 1AKE) as the canonical conformational stress test, and reorganizes the DAG around collective-motion interpretation at 37.3 bits gain. Iteration 3 searches inside the enlarged regime on a mixed validation slice and consolidates the model into the compact multiplicative law in Eq. 7, at 54.3 bits gain. The point of the trajectory is not that successive regressions improved monotonically; it is that the admissible symbolic vocabulary was repeatedly stress-tested, revised, and compressed until the surviving law expressed a new mechanics relation.

Equation 7, with the numerical values above, is the final accepted symbolic world model for the run. It has a clear mechanical reading. The factor $\phi_{pi} = z_p(\log(C_{pi} + \epsilon))$ is a compressed local-compliance coordinate: it is positive for residues whose contact-network environment predicts above-average fluctuation and negative for mechanically buried or strongly constrained residues. The factor $\psi_{pi} = [z_p(|u_{pi2}|) + 2.2678]_+$ is a nonnegative participation weight for the slowest collective mode; the threshold is near the lowest observed value of $z_p(|u_{pi2}|)$ in the mixed validation stage, so the ReLU mostly shifts first-mode participation onto a positive scale and suppresses residues that barely participate in the dominant collective deformation. Their product says that experimental B-factor variation is best compressed by *mode-conditioned compliance*: a residue has high predicted B-factor when it is locally soft in the GNM sense and participates strongly in the dominant collective motion. Local softness that is not aligned with the slow mode is down-weighted, and slow-mode participation without local compliance is insufficient.

The result is not a first-principles derivation from an atomistic Hamiltonian or a crystallographic refinement model, and it is not a neural-network predictor or unconstrained curve fit. The physics sits in the typed compositional pipeline

$$\{\mathbf{r}_{pi}\} \longmapsto A_p \longmapsto \Gamma_p \longmapsto \{(\lambda_{pk}, \mathbf{u}_{pk})\}_k \longmapsto (C_{pi}, |u_{pi2}|) \longrightarrow \widehat{B}_{pi}^{(z)}.$$

The discovery system searches over symbolic compositions of these physically meaningful artifacts and accepts a revised composition only when it compresses broader evidence better after paying for complexity. Thus the learned law is a mechanics-based constitutive surrogate: not "B-factor equals an arbitrary fitted expression," but

$$\text{within-chain B-factor pattern} \sim \text{all-mode elastic compliance} \times \text{slow collective-mode participation}.$$

The log transform is also physically meaningful because raw GNM fluctuations are heavy-tailed; the model discovered that experimental B-factors are better described by a compressed fluctuation scale than by unbounded raw compliance. The key scientific claim is therefore structural: experimental protein flexibility is not governed by local elastic compliance alone, but by local compliance expressed through participation in the dominant collective mode of the contact-network spectrum.

Move	Break type	Generator-reachable new types	Composite-reachable new types	Retracted types	ΔL_M	MDL gain
$0 \to 1$	regime split	ReLU compliance; terminal exposure	boundary product	old linear parameters	+39.1 bits	+9.0 bits
$1 \to 2$	ontology break	none beyond shared GNM base	none beyond shared GNM base	terminal and boundary feature family; old parameters	−14.4 bits	+37.3 bits
$2 \to 3$	regime split	log-compliance; ReLU mode amplitude	mode-conditioned compliance	additive-model parameters	−10.3 bits	+54.3 bits

Table 2: Two-level Kan audit of the Builder/Breaker transitions. Generator-reachable types receive an immediate unary morphism from old schema objects. Composite-reachable types require new intermediate structure or a newly admitted multi-input composition. Parameter objects are singletons in the audit; they are listed only when they affect the interpretation of a transition. The signed ΔL_{model} records whether the symbolic model code grew or shrank. MDL gains are acceptance gains from paired refitting on the same accumulated evidence, not a direct numerical discovery cost. The retracted-types column lists model commitments removed at each transition; retraction is recorded as supersession, so the underlying evidence and its provenance remain available. Read together with ΔL_{model} and the MDL gain, the later accepted transitions reduce the model code length (negative ΔL_{model}) while still yielding the largest MDL gains, so the symbolic model becomes simpler even as it compresses the broader evidence better.

A post-hoc Kan-transport audit of the run makes the categorical content of this discovery explicit (Fig. 4 and Table 2). The audit constructs a finite schema for each accepted outer iteration, transports the old accumulated artifact state into the next schema, and separates three cases. A new type is *generator-reachable* when it receives an immediate unary generating morphism from an old type. A new type is *composite-reachable* when it is not generator-reachable but becomes reachable through new intermediate types or multi-input operations admitted by the enlarged regime. A type is *isolated* when it is unreachable even by such composites. This distinction matters for the final protein law. The features `LogNormCompliance` and `ReLUModeAmpl` are generator-reachable from old physics-derived quantities: they are new transformations of already available compliance and slow-mode amplitude. By contrast, `ModeConditionedCompliance` is composite-reachable only: it appears when the new regime admits the product operation

$$\texttt{LogNormCompliance} \times \texttt{ReLUModeAmpl} \longrightarrow \texttt{ModeConditionedCompliance}.$$

Thus the final discovery is not that GNM compliance or the slow collective mode exists; both are already present in the old schema. The new scientific commitment is the interaction type that lets local compliance be conditioned by participation in a collective deformation, and the MDL gate accepts this commitment with a 54.3-bit gain on the accumulated evidence. The signed model-code changes reinforce the same interpretation: the first accepted transition increases model description length by 39.1 bits, whereas the later accepted transitions reduce it by 14.4 and 10.3 bits. Discovery in this run is therefore not monotone accumulation of terms; it includes retraction and compression of the symbolic vocabulary.

The non-monotonic R^2 trajectory, $0.48 \to 0.68 \to 0.54 \to 0.41$, should therefore be read as evidence-set widening rather than as a failed optimization trace. The four values are not measurements of the same model class on a fixed benchmark. They are descriptive fits on successively enlarged and more adversarial accumulated evidence sets, moving from compact proteins to terminal-flexibility and hinge/domain-motion stress tests and finally to the mixed validation slice. In such a setting, monotone R^2 would be the wrong success criterion: it would reward adding terms to chase heterogeneous data. The MDL gate instead asks whether a revised symbolic structure survives paired comparison on the same enlarged evidence after paying its model-code cost. The final transition is therefore successful not because it maximizes R^2, but because the multiplicative mode-conditioned-compliance law remains the accepted compressed structure after additive and boundary terms fail to pay for themselves.

The inner symbolic search makes this selectivity visible (Fig. 6): in that iteration only 16 of 144 proposed DAG edits survive the MDL gate, and accepted moves include removals as well as additions. Tracking each feature slot across the run makes the churn explicit (Fig. 8): the early iterations only accrete features, whereas iteration 3 explores four slots and retracts three, collapsing to the single surviving mode-conditioned

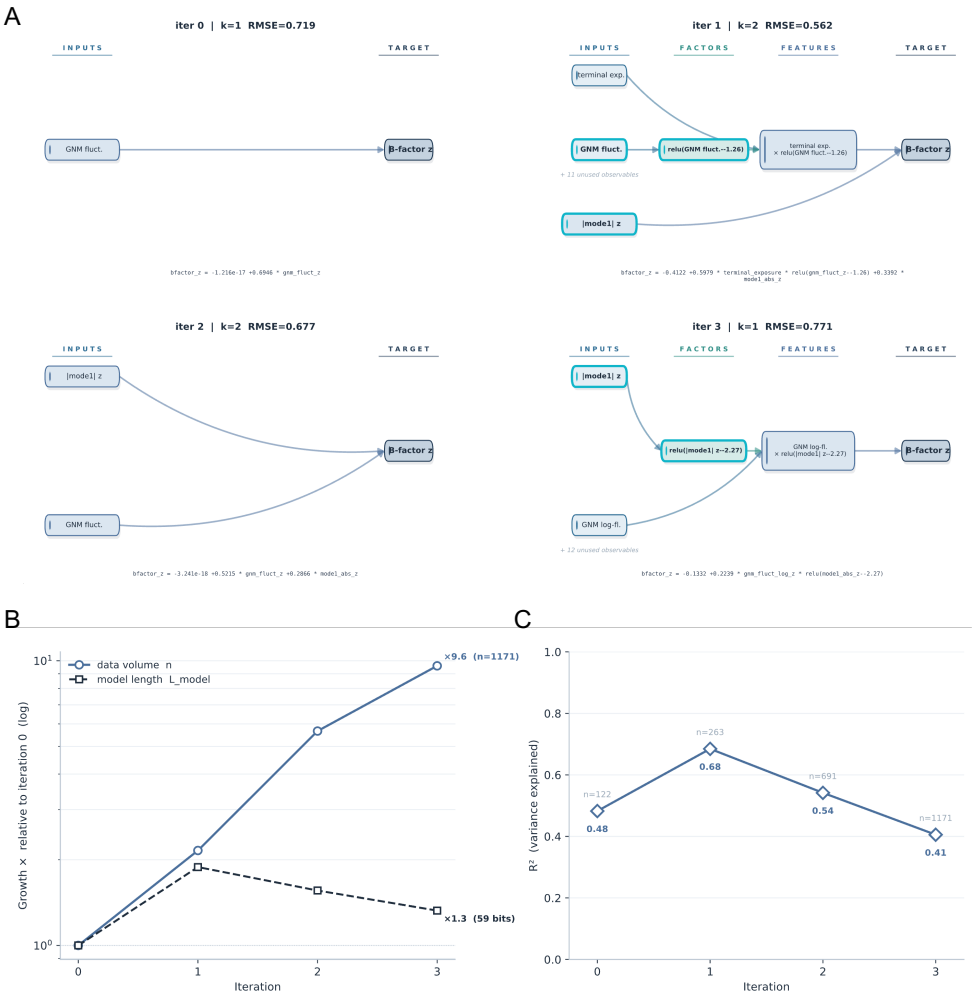
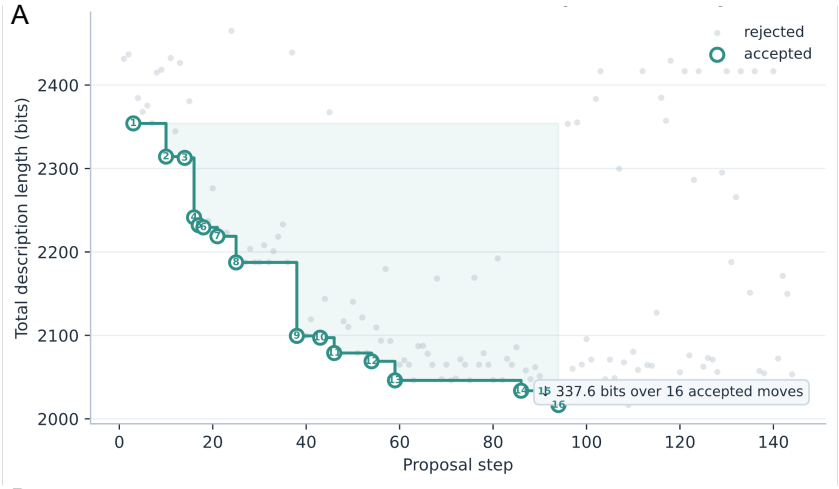


Figure 5: Parsimonious scaling of the discovered world model. (A) Evolution of the world-model DAG across discovery iterations (0–3), read left to right through four stages: inputs (observables), factors (nonlinear transforms, e.g. thresholded ReLU terms), features (terms entering the linear predictor), and the target (B-factor, z-scored). Nodes are colored by stage, edges tinted by source, and nodes new to an iteration are outlined in cyan; each panel lists the iteration, active-feature count k, and RMSE. Complexity peaks at iteration 1 ($k = 2$) and is pruned back to $k = 1$ by iteration 3 as newly revealed slices no longer justify the added term under MDL. (B) Growth relative to iteration 0 (log axis): the data volume n rises $9.6\times$ ($122 \to 1171$ observations) while the model length L_{model} rises only $1.3\times$ ($44 \to 59$ bits, $k \leq 2$)—an order of magnitude more data absorbed without added complexity. (C) Descriptive fit accuracy R^2 versus iteration, with the accumulated sample size n annotated. These values should not be read as a monotone benchmark curve on a fixed test set: each point is evaluated on the evidence available at that stage, and later stages include harder stress-test proteins. The relevant trend is therefore joint parsimony under expanding evidence: n grows $\sim 10\times$ while L_{model} grows only $1.3\times$, and the accepted symbolic law remains compact rather than being allowed to accumulate ad hoc terms. Data based on earlier results [10], with expanded analysis.

law. Aggregated across all four iterations, the gate admits only 25 of 388 proposals (6.4%), with operator-dependent selectivity: structure-recombining moves survive most often (seeds 21%, swaps 11%) while bare feature additions rarely do (3%), and feature removals are a high-yield operator rather than rare cleanup (Fig. 7). The gate is therefore not decorative: it determines which proposed structures become committed scientific artifacts.

A



B

#	STEP	ACCEPTED MOVE	BITS	Δ
1	3	seed add [contact_degree_raw<=6]	2354.1	–
2	10	threshold 6 -> 9.83 on IndLE(contact_degree_raw)	2314.2	-39.8
3	14	add feature res_index_norm * relu(contact_degree_raw-18.8)	2312.7	-1.5
4	16	swap [contact_degree_raw<=9.83] -> contact_degree_z^3	2241.4	-71.3
5	17	swap res_index_norm -> contact_degree_raw^2	2232.0	-9.4
6	18	add feature relu(contact_degree_z-1.73)	2229.4	-2.6
7	21	swap relu(contact_degree_raw-18.8) -> contact_degree_z^4	2218.8	-10.7
8	25	remove feature relu(contact_degree_z-1.73)	2187.4	-31.3
9	38	swap contact_degree_z^3 -> gnm_fluct_log_z	2099.4	-88.0
10	43	swap contact_degree_z^4 -> mode1_abs_z	2097.2	-2.2
11	46	swap contact_degree_raw^2 -> 1	2078.8	-18.4
12	54	add feature gnm_fluct_z * relu(mode1_abs_z--1.75)	2068.9	-9.9
13	59	remove feature mode1_abs_z	2046.0	-22.9
14	86	swap gnm_fluct_z -> gnm_fluct_log_z	2033.6	-12.4
15	91	threshold -1.75 -> -2.27 on ReLU(mode1_abs_z)	2032.6	-1.1
16	94	remove feature gnm_fluct_log_z	2016.5	-16.1

Figure 6: Inner MDL-guided search within a discovery iteration (data from [10]). (A) Hill-climb frontier: total description length (bits) versus proposal step. Faint grey points are rejected proposals; the stepped teal curve is the best-so-far frontier through the accepted moves (numbered markers), which together reduce the description length by 337.6 bits over 16 accepted moves (from 2354.1 to 2016.5 bits). (B) Ledger of the accepted moves, keyed to the numbered markers in (A): each row gives the proposal step, the structural edit (seed, add/remove a feature, swap one feature for another, or adjust a threshold), the resulting total description length in bits, and the change Δ relative to the previous accepted state (negative = improvement). Together the panels show that the model is assembled through many small, individually bit-reducing edits (adding, recombining, thresholding, and pruning features) rather than a single large step.

2.6 CategoryScienceClaw adds a categorical layer to ScienceClaw

CategoryScienceClaw is the second case framework in this paper, and is a wrapper around ScienceClaw. ScienceClaw is the agentic execution substrate where it organizes scientific work as skill-mediated production of artifacts, keeps metadata and parent lineage for those artifacts, exposes unresolved questions as shared open needs, uses pressure and feedback signals to decide what agents should attempt next, and allows active workflows to mutate as evidence changes. In the ScienceClaw $\times$ Infinite architecture, Infinite adds the communication substrate (Fig. 9): structured posts, hypotheses, methods, findings, claim links, votes,

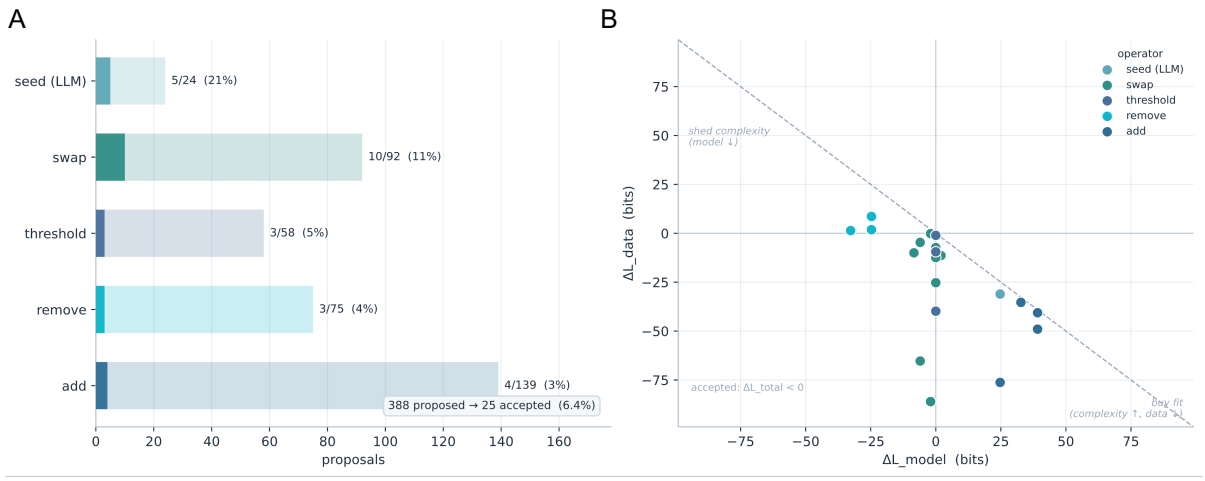


Figure 7: Anatomy of the MDL gate across the discovery run (data from [10]). (A) Gate selectivity by proposal operator: the number of proposals accepted versus proposed, aggregated over all iterations. Of 388 proposals only 25 are accepted (6.4%), and the acceptance rate is strongly operator-dependent (structure-recombining moves survive most often (seed 21%, swap 11%) while bare feature additions rarely do (add 3%)). (B) The minimum-description-length trade-off of each accepted move, shown as its change in model code length ΔL_{model} versus residual data code length ΔL_{data}; deltas are computed within an iteration (between consecutive accepted states, where the evidence set is fixed), giving 21 of the 25 accepted moves. All accepted moves fall below the $\Delta L_{\text{total}} = 0$ frontier (dashed). Operators occupy distinct regions: additions and seeds pay model bits to buy data bits (lower right; mean $\Delta L_{\text{model}} = +33.9$, $\Delta L_{\text{data}} = -50.3$ for additions), threshold tuning improves fit at essentially zero model cost (near-vertical), swaps give model-neutral fit gains (the most frequent accepted move), and removals shed model complexity at a small data-fit cost (upper left; mean $\Delta L_{\text{model}} = -27.4$, $\Delta L_{\text{data}} = +3.9$). Removals are thus a high-yield operator rather than rare cleanup, showing that accepted discovery includes retraction and compression, not only accumulation.

comments, reputation, and moderation. This makes the combined system not merely a tool-calling loop but an artifact-centered scientific workflow and discourse system whose computation and communication traces can both be audited.

CategoryScienceClaw adds the explicit categorical and proof-carrying layer to that substrate. Skills become morphism signatures, artifacts become typed objects with content hashes and parents, open needs become typed holes to be filled by compatible morphisms, worker heartbeats become decentralized reactions, and certificates/audits check type and provenance validity. The layer is domain-general; mechanics is the worked scientific example used here because its model choices, gates, and stress tests can be displayed compactly. In the mechanics run studied here, this layer records mechanics questions, typed computational inputs, candidate model sets, accepted and rejected model artifacts, gate records, stress tests, regime-transition audits, and synthesized figure/report artifacts. Thus the scientific object is not only a final plot or a fluent written claim; it is a typed discovery graph whose morphisms connect scientific inputs to gate-checked interpretations.

The implemented structure should be read precisely. Let A_t be the finite set of ScienceClaw artifacts visible at time t, and let

$$\tau : A_t \longrightarrow T$$

assign each artifact its recorded artifact type. At minimum, this gives a family

$$I_t(X) = \{a \in A_t : \tau(a) = X\}, \qquad X \in T,$$

that is, a copresheaf over the discrete category of implemented artifact types. Each artifact a also records a skill label $\sigma(a)$, a producer agent, a content hash, and a parent list $p(a) = (a_1, \ldots, a_k)$. Thus the realized computational record is not merely a set of files but a typed acyclic hypergraph whose generating operations have the form

$$(a_1, \ldots, a_k) \xrightarrow{\sigma(a)} a.$$

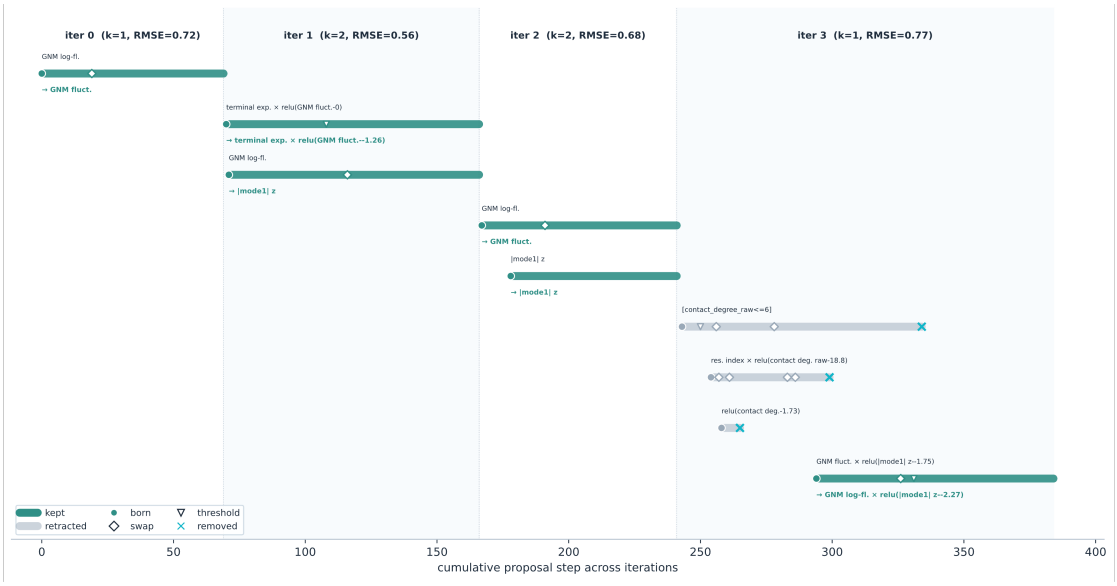


Figure 8: Feature lifecycle across the discovery run, computed from the accepted moves of each iteration's inner search. Each horizontal bar is a feature slot, with markers for its introduction (born, by an add or seed move), factor swaps, threshold tunings, and removal. Slots present at the end of an iteration are kept (teal); slots removed before then are retracted (grey), with the model commitment dropped while its evidence is preserved. The label below each kept slot gives its final symbolic form. The horizontal axis is the cumulative proposal step across iterations; bands give the iteration index, active-feature count k, and RMSE. Iterations 0–2 only accrete features, whereas iteration 3 explores four slots and retracts three, collapsing to the single surviving mode-conditioned law ($k = 1$), so the accepted model is reached by pruning and compression rather than monotone accumulation. Thresholded factors are shown in the internal $\mathrm{relu}(x - \tau)$ form, with fitted $\tau = -\theta$ relative to the $[z + \theta]_+$ parametrization of Eq. 7.

The unary shadow of this hypergraph generates a free provenance category Prov_t; the full multi-parent synthesis structure is more accurately read as a typed multicategory or colored-operadic provenance record. This is weaker than a software-enforced schema category, and should not be overclaimed. It is nevertheless already enough to support the central categorical interpretation: realized scientific work is a typed population of artifacts together with composable, auditable generating operations.

Open needs add the missing-hole structure. A need signal specifies a desired artifact type, query, rationale, and optional preferred skills. In categorical terms, it marks an unfilled target object or missing cone in the active provenance diagram. The ArtifactReactor does not yet solve a formal Kan-extension or lifting problem; it performs the implemented engineering analogue, using pressure scores and schema overlap to find artifacts and skills that may complete the diagram. Mutation then changes the active subgraph by marking artifacts active, rejected, superseded, or pruned while retaining immutable lineage.

Infinite extends this provenance record into scientific discourse. A post is not just text; it is a typed claim artifact with hypothesis, method, findings, evidence, links, and feedback. Links such as extension, contradiction, replication, or citation are discourse morphisms. Votes, comments, and reputation are verifier signals. The current integration therefore defines an implemented publication map from ScienceClaw artifacts and parent relations into Infinite posts, artifact records, and discourse links. It is not yet a certified functor in software, but it has the functorial shape required for one: computational lineage is carried into a public, inspectable, revisable scientific record.

The fiber-network run makes the scientific and categorical stakes compact. The system must decide whether network mechanics is better represented by a scalar fiber-count descriptor or by an orientation-tensor anisotropic stiffness surrogate. The accepted model, rejected alternative, AIC gate, perturbation stress test, and residual typed artifacts make the decision auditable. In categorical terms, CategoryScienceClaw turns model comparison itself into provenance: the rejected isotropic descriptor remains part of the graph, while the accepted anisotropic mechanics object becomes the committed interpretation.

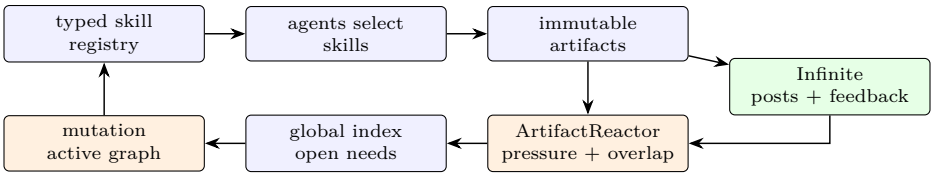


Figure 9: ScienceClaw × Infinite as a distributed typed artifact system. ScienceClaw executes typed skill compositions and records immutable lineage; the ArtifactReactor and mutation layer coordinate active search; Infinite turns computational artifacts into public scientific discourse with feedback that can re-enter the discovery loop.

2.7 CategoryScienceClaw fiber-network mechanics as a typed discovery graph

CategoryScienceClaw makes the categorical framework operational at the level of individual scientific claims while retaining the larger ScienceClaw discovery-system structure of typed skills, immutable lineage, pressure coordination, workflow mutation, and public discourse. The main case here is the fiber-network mechanics run because it gives a compact non-protein example of the categorical layer: a representative network and stress–strain table are transformed into typed descriptors, candidate models, gate records, stress tests, a regime-transition claim, and a final scientific figure/report. The remaining mechanics trees are reported in the separate Supplementary Information file for this version.

The mechanics formalism is explicit. For each fiber orientation θ_i, define

$$n_i = (\cos\theta_i, \sin\theta_i), \qquad A = \frac{\sum_i w_i n_i n_i^T}{\sum_i w_i}.$$

The scalar nematic order parameter and anisotropy ratio are

$$S = \sqrt{\langle\cos 2\theta\rangle^2 + \langle\sin 2\theta\rangle^2}, \qquad \chi = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

The stress–strain surrogate is

$$\sigma = E\epsilon + \sigma_0.$$

For this example, CategoryScienceClaw reports $S = 0.673115$, principal orientation $47.877581°$, stiffness $E = 119.4$ kPa, and $R^2 = 0.999989$ for the linear stress–strain fit. These values make the mechanics interpretation concrete: the system recovers a dominant orientation and a tensile stiffness scale, then tests whether that anisotropic structure explains the response better than fiber count alone.

The candidate-model gate compares an isotropic fiber-count descriptor M_0 with an orientation-tensor anisotropic stiffness surrogate M_1. The declared rule is

$$\text{accept } M_1 \text{ over } M_0 \iff \mathrm{AIC}(M_0) - \mathrm{AIC}(M_1) > 0 \quad \text{and diagnostics pass.}$$

Here M_1 is accepted, M_0 is rejected, and $\Delta\mathrm{AIC} = 123.873782$ (Fig. 10). The scientific claim is therefore not merely that the network has 12 fibers; it is that orientation-tensor structure plus stress–strain stiffness explains the mechanics response better than an isotropic fiber-count descriptor.

Categorically, the run is a typed discovery move. The old computational-input regime can carry files, scalar counts, and stress–strain observations. The enlarged regime contains an orientation tensor, principal axis, anisotropic stiffness surrogate, gate record, and perturbation stress test. In the notation of Definition 6, the objectwise discovery-residual diagnostic is

$$R(A') = I'_{t+1}(A') \setminus \mathrm{im}(\bar{\rho}_{A'}).$$

For the fiber-network run, these objectwise residuals contain the orientation tensor, principal axis, anisotropic stiffness surrogate, gate record, and perturbation stress test. This residual is the part of the accepted state not obtained by simply transporting the old input fibers; it is the new typed mechanics content supplied by the run.

This mechanics gate is close in spirit to the Builder/Breaker MDL gate but differs in what is being charged. Builder/Breaker accepts symbolic DAG revisions when paired refitting reduces total description length on

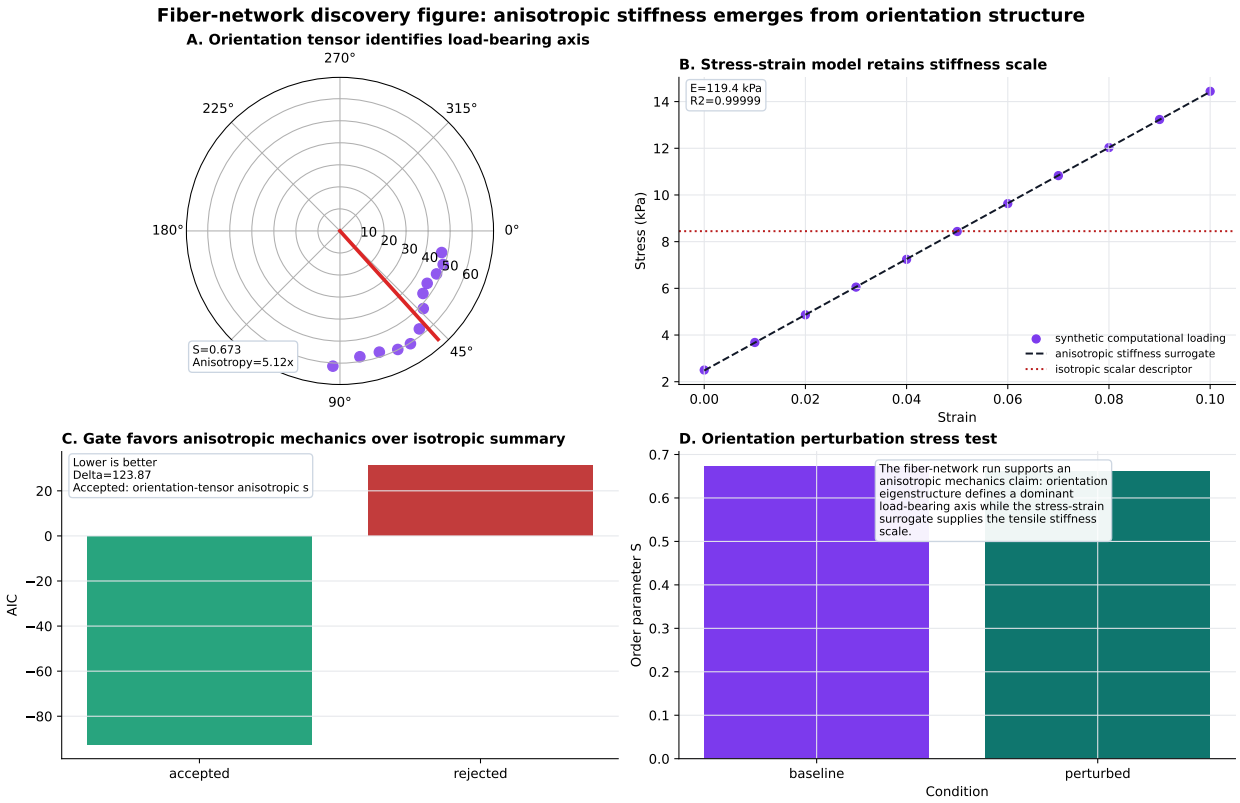


Figure 10: CategoryScienceClaw fiber-network mechanics figure. The figure renders the typed path from a fiber-network mechanics question to typed inputs, candidate models, an accepted orientation-tensor anisotropic stiffness surrogate, a rejected isotropic fiber-count descriptor, an AIC gate, perturbation stress test, regime-transition record, and synthesized scientific report. The result supports anisotropic mechanics: orientation-tensor structure plus stress–strain stiffness explains the response better than fiber count alone.

accumulated protein-mechanics evidence. CategoryScienceClaw accepts the fiber-network model when the richer typed mechanics descriptor clears a model-selection gate and diagnostics on deterministic computational inputs. In both cases rejected alternatives remain first-class audit objects rather than disappearing from the record.

A concrete instance: CategoryScienceClaw fiber-network mechanics. A compressed provenance diagram is

$$a_{\text{net}} \xrightarrow{\text{orientation tensor}} a_A, \qquad a_{\text{stress}} \xrightarrow{\text{linear fit}} a_E, \qquad (a_{\text{net}}, a_{\text{stress}}) \xrightarrow{\text{candidates}} (M_0, M_1),$$
$$(M_0, M_1, a_A, a_E) \xrightarrow{\text{AIC gate}} a_{\text{accepted}}, \qquad (a_{\text{accepted}}, a_{\text{stress test}}) \xrightarrow{\text{synthesize}} a_{\text{figure/report}}.$$

2.8 Instantiations across agentic systems

The same formal pattern appears across several agentic systems developed in this research line, with different regimes, artifact populations, and gates. Table 3 summarizes the mapping. The table is deliberately not a leaderboard. It is a structural comparison of how different systems instantiate typed artifact states, fixed-regime updates, and regime-enlarging mechanisms.

2.9 Relation to existing formalisms

The problem addressed here also has a historical and philosophical lineage. Bacon and Whewell treated scientific method as an organized practice of collecting, comparing, and conceptually ordering phenomena; Peirce emphasized inquiry as a disciplined process for stabilizing belief under doubt [43–45]. Popper, Kuhn,

System	Main artifact regime	Gate or verifier	Discovery-relevant mechanism
ProtAgents [3]	protein sequences, structures, force and dynamics artifacts	agent critique plus physics and ML tool outputs	physics simulators expose failure modes of generated designs
MARS / graph agents [41, 42]	layered material knowledge graphs and substitution candidates	feasibility, manufacturability, and cross-source consistency	cross-layer composition of evidence from heterogeneous corpora
Sparks / SciAgents [4, 5]	research goals, hypotheses, code, results, and reports	novelty, interestingness, executable results	ideation-to-experiment loops over a research-artifact schema
Builder/Breaker [10]	protein-mechanics evidence and symbolic DAG world models	paired MDL compression	adversarial evidence forces retractions and regime enlargement
ScienceClaw × Infinite with CategoryScienceClaw layer [9]	execution artifacts and open needs; discourse posts, claim links, feedback; categorical objects, morphisms, lineage, gates, figures, and reports	pressure scoring, schema overlap, mutation status, discourse feedback, reputation, and proof certificates	plannerless artifact exchange plus public claim verification, lifted into typed categorical provenance

Table 3: Agentic systems as typed artifact systems. Each system has a schema, artifact population, gate, and mechanism by which the active regime can be stressed or enlarged.

and Lakatos then made failure, anomaly, paradigm change, and research-programme revision central to accounts of scientific progress [13–15].

Goethe and Whitehead supply a complementary lineage in which form, relation, and process are primary scientific objects rather than incidental descriptions [28, 46]. Polanyi and Hacking are also relevant: scientific knowledge is not only explicit proposition but skillful practice, and experimental intervention is part of what makes a representation scientifically real [47, 48]. The contribution of the present paper is to give this broad intuition an auditable mathematical substrate for AI systems: typed artifacts, composable provenance, explicit gates, and verified regime transitions.

2.9.1 Relation to scientific knowledge graphs and workflow provenance

Scientific knowledge graphs and workflow-provenance models already provide important ways to represent entities, relations, computational activities, and data products in a graph [49–52]. The present contribution is therefore not the observation that computation can appear in a graph. Rather, it is to give a categorical semantics in which knowledge, computation, verification, rejection, public discourse, and schema-changing discovery are components of a single scientific state (Definitions 1, 2, 3, and 6). Computation is represented both at the type level, as a morphism $f : A \to B$, and at the artifact level, as a realized map $I_t(f) : I_t(A) \to I_t(B)$ or, in the multicategorical case, as a multi-parent operation (Proposition 8). Discovery is then not graph completion inside a fixed ontology, but a verified regime transition together with residual content beyond functorial transport (Proposition 4).

2.9.2 Relation to categorical learning, coalgebra, and model selection

The framework is adjacent to, but distinct from, categorical deep learning. Categorical accounts of neural networks and learning, including backpropagation as a functor and categorical deep learning programs based on parametric maps and lenses, clarify the structure of model architectures and training procedures [53–56]. The present paper instead studies systems whose state is a typed scientific artifact population and whose most important operation may be a change of regime. The unit of analysis is therefore not only a trained model, but a scientific workflow capable of altering its own schema of admissible artifacts.

The framework also touches coalgebra and dynamical systems. A fixed-regime update resembles a state transition system and can be studied with coalgebraic ideas such as bisimulation and final behavior [57, 58]. The difference is that discovery changes the state space itself: the relevant trajectory lives over a base of regimes rather than inside one fixed carrier. This is why the indexed or fibered picture is natural.

We finally note that the gate connects the framework to MDL, Bayesian model selection, algorithmic information, and open-endedness. MDL and Solomonoff-style compression provide one rigorous account

Aspect	Knowledge graph / workflow-provenance view	Categorical knowledge–computation graph
Vocabulary	Fixed or slowly curated ontology of entity and relation types	Schema category $\mathcal{S}_b$ of artifact types and admissible operations, with possible transition $u : \mathcal{S}_b \to \mathcal{S}_{b'}$
Computation	Activities, services, scripts, or workflow steps that consume inputs and generate outputs	Typed morphisms $f : A \to B$ and realized artifact maps $I_t(f) : I_t(A) \to I_t(B)$, with multi-parent synthesis represented by a typed multicategory or colored-operadic provenance object
Scientific commitment	Validity is often represented as annotations, confidence/provenance metadata, or downstream checks	Commitment, rejection, supersession, and stress testing are represented by gates $\mathcal{V}_b$, description-length or model-selection functionals L_b, diagnostics, and retained rejected artifacts
Public discourse	Claims, citations, comments, and replications may be represented as graph nodes or external communication records	A discourse layer D_t and publication map π_t carry audited artifact paths or hyperpaths into claim, evidence, objection, replication, and validation objects
Dynamics	Graph completion, link prediction, or workflow execution inside a fixed representational frame	Fixed-regime search is Φ_b; discovery is a verified transition with transported evidence $\mathrm{Lan}_u I_t$ and comparison map $\bar{\rho} : \mathrm{Lan}_u I_t \to I'_{t+1}$
Discovery content	New facts, links, generated outputs, or completed workflow products	Objectwise residual artifacts $I'_{t+1}(A') \setminus \mathrm{im}(\bar{\rho}_{A'})$, together with recorded regime-level additions such as new object types, morphisms, tools, verifiers, or grammar productions needed to express them

Table 4: From knowledge graphs and workflow provenance to categorical knowledge–computation graphs.

of why a simpler world model that explains broader evidence is preferable [59–63]. Open-endedness and automated-scientist systems study how systems escape fixed objectives or generate new scientific tasks [6, 7, 64, 65]. Scientific machine learning supplies many of the typed operators that appear as morphisms in our schemas, from physics-informed learning to operator learning and equation-free multiscale computation (e.g., [66–69]). Similarly, the sparse identification of governing equations [70] and the extraction of symbolic models via inductive biases [71] provide established paradigms for discovering interpretable physical laws, which appear in our framework as composable, complexity-penalized morphisms. The contribution here is to place these components into one typed, provenance-preserving, regime-changing account of discovery.

2.10 Design principles and open problems

The framework suggests five open problems articulated as follows:

Open Problem (Convergence on growing regimes)**.** Classical fixed-point theory assumes a fixed underlying space. Discovery systems iterate Φ_b while also producing regime transitions $b_0 \to b_1 \to b_2 \to \cdots$. Under what conditions does the sequence of transported artifact states converge to a stable object in an appropriate colimit or indexed category? When is non-convergence productive exploration, and when is it unproductive oscillation?

Open Problem (Scaling laws for discovery)**.** Standard scaling laws measure loss or benchmark performance inside a fixed regime. Discovery requires a different observable: the rate, quality, and accepted value of regime enlargements. How do discovery rates scale with model size, tool diversity, simulator fidelity, evidence heterogeneity, schema richness, and verifier strength?

Open Problem (Verification tooling for agentic loops)**.** The framework requires provenance preservation and gate compatibility. Practical systems need tooling that can replay downstream artifact chains, check approximate naturality, account for description length or pressure scores, and record retractions without deleting provenance.

Open Problem (Learning the base schema category)**.** The framework assumes a schema category $\mathcal{S}_b$. Current systems construct it by engineering choice. A major open problem is to learn useful schema categories from corpora, tool signatures, code, figures, equations, and laboratory protocols while keeping morphisms scientifically meaningful. Functorial data migration, olog-style modeling, and category-theoretic accounts of hierarchical materials and building-block replacement provide a starting toolkit [20, 23–25]; the open problem is computational: how to learn $\mathcal{S}_b$ and an accompanying description-length functional L_b from data at the scale of working scientific corpora.

Open Problem (Multicategorical discovery)**.** The Kan audit in Section 2.5 distinguishes generator-level transport from composite reachability through newly admitted multi-input morphisms. Proposition 8 gives the corresponding multicategorical obstruction under the standard operadic Kan-extension setup. The open problem is now computational and statistical: how can an agentic system learn the typed multicategory or colored-operadic schema M_b from traces, tool signatures, equations, figures, and artifact lineages, and how can it estimate the artifact and operation components of discovery cost at scale?

3 Conclusions and Outlook

This paper has argued that agentic scientific discovery requires two levels of structure. Inside a fixed regime, an agentic system updates typed artifact populations: it adds data, models, simulations, hypotheses, critiques, and reports while preserving provenance. At that level, copresheaves, categories of elements, natural transformations, and endofunctorial dynamics give a precise language for structures already present in working discovery systems. Discovery in the stronger sense occurs when evidence forces a transition to a new regime, and left Kan extension gives the mathematical language for transporting old artifacts into the enlarged vocabulary.

The Builder/Breaker protein-mechanics case shows this structure quantitatively: a symbolic world model is revised by adversarial evidence and an MDL gate, with rejected edits and retractions visible in the audit trail. CategoryScienceClaw shows the same structure as a proof-carrying categorical layer over the ScienceClaw $\times$ Infinite substrate: typed skills, immutable lineage, open needs, pressure-based coordination, and public discourse become typed objects and morphisms, and a fiber-network mechanics run records its accepted model, rejected alternative, AIC gate, and stress test as inspectable provenance.

More broadly, the work suggests a reciprocal program: AI can accelerate mechanics, but mechanics can also discipline AI by providing concepts of state, load, failure, invariance, admissible transformation, and constitutive closure for self-revising discovery systems.

CategoryScienceClaw is already close to the categorical picture because it makes the ScienceClaw data model and workflow discipline explicit. It has typed artifact metadata, skill labels, immutable parent lineage, content hashes, open needs, pressure-ranked reactions, mutation of active status, public discourse objects with feedback signals, and domain-specific gate records. The categorical layer adds object and morphism signatures, typed needs, proof certificates, and audit/replay checks. This is enough to read the current implementation as a typed artifact family equipped with an acyclic provenance hypergraph and a publication map into a discourse graph. The mechanics case in this paper is one scientific audit within that broader substrate: accepted and rejected model objects, stress tests, regime-transition residuals, and figure/report artifacts are committed as inspectable elements of the graph.

The next step is therefore not to replace ScienceClaw, but to lift structures already present in ScienceClaw and CategoryScienceClaw into explicit mathematical objects. Skill manifests would become morphism signatures with declared input objects, output objects, constraints, and verifiers. Artifact stores would become materialized fibers of a copresheaf over that schema. Multi-parent synthesis would be treated as a typed multicategory or operadic composition, rather than only as a parent list. Open needs would become typed holes or lifting problems. Public discourse would become a category whose objects are claims, posts, artifacts, evidence bundles, objections, replications, and validation signals. The CategoryScienceClaw publication map would then be a checked map from provenance to discourse, preserving identities, parent-child composition, and validation status.

Making this explicit would change the operational status of the platform. The system would no longer merely display provenance; it could verify that provenance diagrams commute, that every public claim has an admissible artifact path, that retractions and supersessions preserve old evidence, and that new artifact types are introduced through recorded schema transitions. In the language of this paper, ScienceClaw equipped

with the CategoryScienceClaw layer would move from an operational typed-artifact platform to an executable categorical discovery substrate. The impact would be practical: stronger audit trails, machine-checkable claim lineage, better comparison among autonomous investigations, and a route toward measuring discovery as the residual content added beyond transport of prior evidence.

The central insight is therefore that search is iteration inside a typed scientific regime; discovery is a verified regime transition—enlargement, restructuring, or compression—together with the residual content that lies outside functorial transport of prior evidence. This statement is mathematical enough to constrain future theory and concrete enough to guide the design of future AI discovery systems.

4 Materials and Methods

4.1 Regimes, schemas, and copresheaves

Definition 1 (Discovery regime)**.** A **discovery regime** is a tuple

$$b = (\mathcal{S}_b, \Gamma_b, \mathcal{V}_b, L_b),$$

where $\mathcal{S}_b$ is a small schema category of artifact types and allowed operations, Γ_b is a grammar for composing admissible artifacts or model structures, $\mathcal{V}_b$ is a verifier or gate, and L_b is an optional description-length functional. The functional L_b may be presented either absolutely on artifact states or as a relative (conditional) length functional; we write $L_b(I \mid J)$ for the code length of I given a subobject $J \hookrightarrow I$, with $L_b(I) \equiv L_b(I \mid \varnothing)$ recovering the absolute case. In this paper, $\mathcal{S}_b$ carries the main type-and-operation structure; Γ_b, $\mathcal{V}_b$, and L_b record the additional grammar, acceptance, and scoring data attached to that structure. When these additional data are clear from context, "regime" is used informally as shorthand for the schema and its associated commitments.

Definition 2 (Artifact-state copresheaf)**.** For a fixed regime b, an artifact state at time t is a copresheaf

$$I_t : \mathcal{S}_b \to \mathbf{Set}.$$

For each object $A \in \mathrm{Obj}(\mathcal{S}_b)$, $I_t(A)$ is the set of artifacts of type A. For each morphism $f : A \to B$, $I_t(f) : I_t(A) \to I_t(B)$ maps artifacts along an allowed operation.

Definition 3 (Category of elements)**.** The realized provenance category of I_t is the category of elements $\int_{\mathcal{S}_b} I_t$. Its objects are pairs (A, x) with $x \in I_t(A)$. A morphism $(A, x) \to (B, y)$ is a morphism $f : A \to B$ in $\mathcal{S}_b$ such that $I_t(f)(x) = y$.

For partially realized scientific workflows, $I_t(f)$ may be a partial function, relation, or span rather than a total function. The total-function presentation is the clean base case; partiality can be handled by replacing **Set** with a category of partial maps, relations, spans, or typed records with status events.

4.2 Fixed-regime update and endofunctoriality

Let $\mathcal{K}_b$ be a category whose objects are artifact-state copresheaves over $\mathcal{S}_b$ and whose morphisms are provenance-preserving refinements. In the strict presentation used for the structural observations and the Kan proposition, a morphism $\alpha : I_t \to J_t$ is a componentwise injective natural transformation whose components $\alpha_A : I_t(A) \to J_t(A)$ preserve artifact identity, type, and provenance metadata, up to the equivalence or tolerance declared by $\mathcal{V}_b$. The injectivity assumption is the mathematical version of a practical audit rule: a refinement may add, annotate, or supersede artifacts, but it may not silently identify two previously distinct accepted artifacts.

Definition 4 (Fixed-regime update)**.** A **fixed-regime agentic update** is an endomap on artifact states,

$$\Phi_b : \mathrm{Obj}(\mathcal{K}_b) \to \mathrm{Obj}(\mathcal{K}_b),$$

that becomes an endofunctor $\Phi_b : \mathcal{K}_b \to \mathcal{K}_b$ when it maps refinement morphisms to refinement morphisms and preserves identities and composition.

A realized committed trajectory will be written as a sequence

$$I_t \xrightarrow{\delta_t} I_{t+1} = \Phi_b(I_t),$$

where δ_t is the refinement morphism that embeds the previous committed ledger into the next committed ledger. Thus Φ_b acts on states and on refinements between alternative states, whereas δ_t is the particular accepted step in one realized run.

The update may be implemented by agents, simulators, retrieval systems, symbolic search, or human feedback. The categorical requirement is not that the implementation be deterministic or neural-free, but that the committed state preserve typed provenance. Concretely, an implementation earns the endofunctor notation only when each refinement $\alpha : I \to J$ induces a refinement $\Phi_b(\alpha) : \Phi_b(I) \to \Phi_b(J)$ and these induced maps respect identity refinements and composition of refinements. In stochastic or asynchronous implementations, this condition is imposed on the committed-state relation, stochastic kernel, or Kleisli morphism rather than on every raw execution trace.

4.3 Builder, Breaker, and gates

Definition 5 (Builder/Breaker system)**.** Within a regime b, a Builder/Breaker system consists of a proposal mechanism B, an evidence-producing mechanism K, and a gate $\mathcal{V}_b$. The Breaker K selects or generates evidence intended to stress the current artifact state. The Builder B proposes candidate increments or revisions. The gate $\mathcal{V}_b$ decides whether the candidate is committed, rejected, superseded, or held for review.

For the protein-mechanics case, the gate is MDL:

$$\mathcal{V}_b(M', M; D) = 1 \quad \Longleftrightarrow \quad L_{\text{model}}(M') + L_{\text{data}}(D \mid M') < L_{\text{model}}(M) + L_{\text{data}}(D \mid M). \tag{8}$$

Comparisons are paired: both models are refit on the same accumulated data D. For CategoryScienceClaw, $\mathcal{V}_b$ combines mechanics-specific gates with platform-level commitment signals: candidate models are compared by model-selection scores and retained only when diagnostics pass, while pressure, mutation status, discourse feedback, and reputation help determine which artifacts remain active in the broader discovery graph.

4.4 Verified regime transition and Kan transport

Definition 6 (Verified regime transition)**.** Given an old state $I_t \in \mathcal{K}_b$ and a new accepted state $I'_{t+1} \in \mathcal{K}_{b'}$, a **verified regime transition** from b to b' relative to (I_t, I'_{t+1}) consists of a schema functor

$$u : \mathcal{S}_b \to \mathcal{S}_{b'}$$

together with a preservation natural transformation

$$\rho : I_t \longrightarrow u^* I'_{t+1},$$

where $u^* : [\mathcal{S}_{b'}, \mathbf{Set}] \to [\mathcal{S}_b, \mathbf{Set}]$ is restriction along u. The state-dependence is part of the data: a verified transition is a piece of structure attached to a specific pair of states, not a universal property of the schemas alone. The transition satisfies four compatibility conditions:

1. u is faithful on morphisms and injective on objects, except where the transition explicitly records a type split, quotient, or recoding;
2. ρ is componentwise injective and provenance-preserving, so old accepted artifacts remain inspectable after the transition;
3. old commitments remain accepted after transport: for predicate-valued gates, $\mathcal{V}_b(I_t) = 1$ implies $\mathcal{V}_{b'}(\text{im}(\bar{\rho})) = 1$, where $\bar{\rho} : \text{Lan}_u I_t \to I'_{t+1}$ is the adjoint transpose of ρ.
4. the new state is itself gate-accepted in its own regime: $\mathcal{V}_{b'}(I'_{t+1}) = 1$.

If a description-length functional is present, the transported old state has no larger code length than the original old state, $L_{b'}(\text{im}(\bar{\rho})) \leq L_b(I_t)$, or else the increase is recorded as an explicit recoding cost. The transition is **nontrivial** when $\mathcal{S}_{b'}$ contains a new object, morphism, verifier, grammar production, or tool class not generated inside $u(\mathcal{S}_b)$, or when I'_{t+1} contains accepted artifacts not in $\text{im}(\bar{\rho})$.

The transported old artifact state is the left Kan extension

$$(\text{Lan}_u I_t)(A') \cong \text{colim}_{(uA \to A') \in (u \downarrow A')} I_t(A), \tag{9}$$

where the colimit exists because **Set** is cocomplete and $\mathcal{S}_b$ is small. If u is fully faithful, the unit $I_t \to u^* \operatorname{Lan}_u I_t$ is an isomorphism, so old fibers are preserved by transport alone. If u is only faithful, if it recodes old object names, or if the new schema adds morphisms among old types, preservation is not automatic; it is supplied by the explicit map ρ in Definition 6. When the comma category $(u \downarrow A')$ is empty—that is, when A' receives no morphisms from $u(\mathcal{S}_b)$—the colimit is the initial object of **Set**, giving $(\operatorname{Lan}_u I_t)(A') = \varnothing$. New evidence then populates such fibers, or activates new morphisms absent from the old regime.

4.5 Structural observations and the Kan obstruction

The first three statements are closure and verification observations: they unpack what is already forced by a fixed schema, provenance-preserving refinement, and a declared gate. The first nontrivial categorical obstruction is the Kan statement that follows, where the comparison map makes residual content unavoidable. The later propositions record the functorial consequences needed for auditability, composition of transitions, and the multicategorical reading of product-like discovery moves.

Observation 1 (Fixed-regime reachability)**.** Let b be fixed and let $\Phi_b : \mathcal{K}_b \to \mathcal{K}_b$ be an endofunctor on artifact states over $\mathcal{S}_b$. For any state I_0 and any $n \in \mathbb{N}$, every object appearing in the provenance category $\int \Phi_b^n(I_0)$ lies over an object of $\mathcal{S}_b$. Thus finite iteration of a fixed-regime update cannot by itself create an artifact of a type outside $\mathcal{S}_b$.

Proof. Each state in $\mathcal{K}_b$ is by definition a copresheaf on $\mathcal{S}_b$. Applying Φ_b returns another object of $\mathcal{K}_b$, hence another copresheaf on the same schema. The category of elements of any such copresheaf has objects (A, x) with $A \in \operatorname{Obj}(\mathcal{S}_b)$. Induction on n gives the claim. □

Observation 2 (Verification requires provenance preservation and a gate)**.** Let $\alpha : I \to J$ be a committed update inside a fixed regime b. If the update is verified in the sense of the regime, then (i) α preserves prior provenance up to the equivalence or tolerance declared by $\mathcal{V}_b$, and (ii) α passes the gate $\mathcal{V}_b$. Conversely, if both conditions hold and $\mathcal{V}_b$ is the regime's declared acceptance criterion, then the update is verified.

Proof. If provenance preservation fails, then at least one previously accepted artifact chain cannot be re-expressed after the update, so the update is not verified. If the gate fails, the update violates the regime's declared commitment criterion. Conversely, provenance preservation keeps old accepted diagrams valid, and the gate supplies the regime-specific acceptance judgment. □

Observation 3 (Nontrivial discovery requires regime-extending structure)**.** Let b be fixed. A nontrivial discovery move that produces an artifact over a type, morphism, verifier, grammar production, or tool class absent from $\mathcal{S}_b$ cannot be obtained by finite iteration of $\Phi_b : \mathcal{K}_b \to \mathcal{K}_b$ alone. It requires a verified regime transition $u : \mathcal{S}_b \to \mathcal{S}_{b'}$ or an equivalent schema-generating mechanism.

Proof. By Observation 1, finite iteration of Φ_b yields artifact states over $\mathcal{S}_b$. Hence it cannot create an artifact whose type is not an object of $\mathcal{S}_b$. Moreover, because $b = (\mathcal{S}_b, \Gamma_b, \mathcal{V}_b, L_b)$ is fixed during iteration, the grammar, verifier, and available tool classes are fixed as well. A move requiring a new type, morphism, verifier, grammar production, or tool class therefore cannot be the result of iteration inside $\mathcal{K}_b$ alone; it requires a transition to a regime in which the new structure exists. □

Proposition 4 (Kan obstruction to transport-only discovery)**.** *Let $u : \mathcal{S}_b \to \mathcal{S}_{b'}$ and $\rho : I_t \to u^* I'_{t+1}$ be a verified regime transition (Definition 6). Let*

$$\bar{\rho} : \operatorname{Lan}_u I_t \longrightarrow I'_{t+1}$$

be the unique adjoint transpose of ρ under $\operatorname{Lan}_u \dashv u^*$. Three statements follow.

1. *If $(u \downarrow A') = \varnothing$, then transport is empty at A':*

$$(\operatorname{Lan}_u I_t)(A') = \varnothing.$$

If $I'_{t+1}(A') \neq \varnothing$, then $\operatorname{im}(\bar{\rho}_{A'}) = \varnothing$ and the residual at A' is forced to be nonempty. Every artifact accepted at such an A' must be supplied from outside transport of the old regime.

2. *The image* $\mathrm{im}(\bar{\rho}) \subseteq I'_{t+1}$ *is a sub-copresheaf of transported evidence. At each object* A'*, the pointwise residual set*

$$\mathcal{R}(A') = I'_{t+1}(A') \setminus \mathrm{im}(\bar{\rho}_{A'})$$

records artifacts at that type not obtained by functorial transport of old evidence. The categorical object is the inclusion $\mathrm{im}(\bar{\rho}) \hookrightarrow I'_{t+1}$*; the complements* $\mathcal{R}(A')$ *are an objectwise engineering diagnostic and need not themselves form a sub-copresheaf.*

3. *When a relative description-length functional* $L_{b'}(- \mid -)$ *is present, the* ***discovery cost*** *of the move is*

$$L_{b'}\big(I'_{t+1} \mid \mathrm{im}(\bar{\rho})\big),$$

the code length of the post-transition state given transported evidence. If $L_{b'}$ *is additive on such inclusions, this can be written as* $L_{b'}(I'_{t+1}) - L_{b'}(\mathrm{im}(\bar{\rho}))$*. In particular, any description-length functional assigning positive conditional length to nonempty forced residuals gives a strictly positive local discovery cost at such a type.*

Proof. The adjunction $\mathrm{Lan}_u \dashv u^*$ gives a natural bijection

$$\mathrm{Nat}(\mathrm{Lan}_u I_t, I'_{t+1}) \cong \mathrm{Nat}(I_t, u^* I'_{t+1}),$$

so ρ determines the unique map $\bar{\rho}$. Statement (i) is the pointwise formula (9) for an empty index category: a colimit over the empty diagram in **Set** is the initial object $\varnothing$. Hence the component of $\bar{\rho}$ at A' has empty domain and empty image, so any accepted element of $I'_{t+1}(A')$ lies outside transport. For (ii), the pointwise image of a natural transformation between Set-valued functors is a subfunctor: naturality sends elements in $\mathrm{im}(\bar{\rho}_{A'})$ to elements in $\mathrm{im}(\bar{\rho}_{B'})$ along every morphism $A' \to B'$. Therefore $\mathrm{im}(\bar{\rho})$ is the transported-evidence substate, and the pointwise complement records artifacts not in that transported image. Statement (iii) is the definition of the relative description-length functional on the inclusion $\mathrm{im}(\bar{\rho}) \hookrightarrow I'_{t+1}$, with strict positivity following from the stated positivity assumption on forced residuals. □

Proposition 5 (Refinements lift to realized provenance)**.** *Let* b *be fixed and let* $\alpha : I \to J$ *be a refinement morphism in* $\mathcal{K}_b$*, represented in the strict presentation by a componentwise injective natural transformation. Then* α *induces a faithful functor*

$$\int \alpha : \int_{\mathcal{S}_b} I \longrightarrow \int_{\mathcal{S}_b} J$$

defined by $(A, x) \mapsto (A, \alpha_A(x))$ *on objects and by sending a realized operation* $f : (A, x) \to (B, y)$ *to the same schema morphism* $f : (A, \alpha_A(x)) \to (B, \alpha_B(y))$*. The assignment* $\alpha \mapsto \int \alpha$ *preserves identities and composition of refinements.*

Proof. If $f : (A, x) \to (B, y)$ is a morphism in $\int_{\mathcal{S}_b} I$, then $I(f)(x) = y$. Naturality of α gives

$$J(f)(\alpha_A(x)) = \alpha_B(I(f)(x)) = \alpha_B(y),$$

so the same schema morphism f is a morphism in $\int_{\mathcal{S}_b} J$. Thus $\int \alpha$ is a functor. It is faithful because it does not change the underlying schema morphism on any hom-set. Identity refinements act as identity functors, and for refinements $\alpha : I \to J$ and $\beta : J \to L$, the componentwise equality $(\beta \circ \alpha)_A = \beta_A \circ \alpha_A$ gives $\int(\beta \circ \alpha) = \int \beta \circ \int \alpha$. □

Remark 2. Proposition 5 is the formal audit rule behind the artifact-ledger language. If a refinement adds, annotates, or supersedes artifacts without identifying previously distinct accepted artifacts, then the realized provenance graph of the predecessor embeds faithfully into the realized provenance graph of the successor. Older lineage queries remain meaningful after accepted refinements.

Proposition 6 (Verified regime transitions compose)**.** *Let* (u, ρ) *be a verified regime transition from* b *to* b' *relative to* (I_t, I'_{t+1})*, and let* (v, σ) *be a verified regime transition from* b' *to* b'' *relative to* (I'_{t+1}, I''_{t+2})*. Assume that gates are hereditary on declared transported substates: whenever a transported state is accepted, any transported substate selected by a preservation map is also accepted. Then*

$$(v \circ u, \tau), \qquad \tau = (u^*\sigma) \circ \rho : I_t \longrightarrow (v \circ u)^* I''_{t+2},$$

is a verified regime transition from b to b'' relative to (I_t, I''_{t+2}). Moreover, under the canonical isomorphism $\mathrm{Lan}_{v\circ u} \cong \mathrm{Lan}_v \mathrm{Lan}_u$, the comparison map $\bar{\tau}$ factors as

$$\mathrm{Lan}_{v\circ u} I_t \cong \mathrm{Lan}_v \mathrm{Lan}_u I_t \xrightarrow{\mathrm{Lan}_v(\bar{\rho})} \mathrm{Lan}_v I'_{t+1} \xrightarrow{\bar{\sigma}} I''_{t+2}.$$

Proof. The composite $v \circ u$ preserves morphism faithfulness and object injectivity except at type splits, quotients, or recodings explicitly recorded by the two transitions. Since restriction along a composite satisfies $(v \circ u)^* = u^* v^*$, the map $\tau = (u^*\sigma) \circ \rho$ is natural, componentwise injective, and provenance-preserving because both factors are. If $\mathcal{V}_b(I_t) = 1$, the first transition accepts the transported image in b', and by the acceptance condition on the new state $\mathcal{V}_{b'}(I'_{t+1}) = 1$. The gate condition of the second transition then gives $\mathcal{V}_{b''}(\mathrm{im}(\bar{\sigma})) = 1$. Since the displayed factorization makes $\mathrm{im}(\bar{\tau}) \subseteq \mathrm{im}(\bar{\sigma})$ a transported substate selected by a preservation map, heredity gives $\mathcal{V}_{b''}(\mathrm{im}(\bar{\tau})) = 1$. The factorization of $\bar{\tau}$ follows from uniqueness of left adjoints to restriction: $\mathrm{Lan}_{v\circ u}$ and $\mathrm{Lan}_v \mathrm{Lan}_u$ are both left adjoint to $u^* v^*$, and the displayed composite is the adjoint transpose of τ. □

Remark 3. This proposition justifies reading a multi-stage investigation as one audited discovery move when every intermediate transition records a preservation map and clears its gate. In the Builder/Breaker case, the four accepted outer iterations can therefore be viewed both locally, transition by transition, and globally, as a composite map transporting the initial evidence into the final symbolic regime while accumulating residual content.

Proposition 7 (Declared old evidence remains inspectable)**.** *Let (u, ρ) be a verified regime transition from b to b' relative to (I_t, I'_{t+1}). For every old type A and artifact $x \in I_t(A)$, the component ρ_A gives a unique preserved artifact $\rho_A(x) \in I'_{t+1}(uA)$ with the same declared identity and provenance metadata, up to the equivalence or tolerance specified by the transition. If $J \hookrightarrow I_t$ is an old accepted evidence substate and the gate-compatibility condition declares its transported image accepted in b', then all artifacts of J remain accepted as transported evidence in I'_{t+1}.*

Proof. Componentwise injectivity and provenance preservation of ρ are part of Definition 6. Therefore distinct old artifacts remain distinct after applying ρ, and their recorded lineage remains inspectable in the new state. For an accepted evidence substate $J \hookrightarrow I_t$, restricting ρ gives a preservation map from J into $u^* I'_{t+1}$, equivalently a comparison map from its transported image into I'_{t+1}. The gate-compatibility hypothesis then says precisely that this transported image is accepted in b'. □

Remark 4. The proposition is deliberately about preserved evidence, not about every old model commitment. Discovery systems may explicitly retract, compress, or supersede symbolic model terms when the new gate rejects them on accumulated evidence. What is prohibited is silent deletion of the old evidential record. In the Builder/Breaker audit, the additive-model parameters retracted at the $2 \to 3$ transition are model commitments recorded as superseded; the protein evidence and its provenance remain available for the accepted multiplicative law.

Proposition 8 (Multicategorical Kan obstruction)**.** *Assume a regime is presented by a small typed multicategory, or equivalently a colored-operadic schema, whose multimorphisms*

$$\phi : (A_1, \ldots, A_k) \longrightarrow B$$

declare admissible multi-input scientific operations. Assume artifact states are Set-valued algebras on these schemas and that restriction along a multifunctor $u : M_b \to M_{b'}$ admits a left adjoint $\mathrm{Lan}^{\mathrm{m}}_u$ computed by the usual operadic comma-colimit construction [72]. For a verified multicategorical transition with preservation map $\rho : I_t \to u^* I'_{t+1}$ and adjoint comparison map

$$\bar{\rho} : \mathrm{Lan}^{\mathrm{m}}_u I_t \longrightarrow I'_{t+1},$$

the following hold.

1. *If the operadic comma category of operations from transported old types into a new type B is empty in every arity, then $(\mathrm{Lan}^{\mathrm{m}}_u I_t)(B) = \varnothing$. Any accepted artifact at B is forced residual content, exactly as in Proposition 4.*

2. *If B has no unary generating morphism from the old image but receives a new k-ary multimorphism from transported or generator-reachable inputs, then B is not isolated in the multicategorical regime. Its transported part is generated by formal composites of transported input artifacts through that multimorphism, while the newly admitted multimorphism itself is residual schema content.*
3. *If the description-length functional separates artifact content from operation-registry content, then the relative discovery cost splits into an artifact residual term and an operation residual term; without additivity, the same statement holds as a subadditive upper bound.*

Proof. The proof is the multicategorical analogue of Proposition 4. By assumption, $\mathrm{Lan}_u^{\mathrm{m}}$ is left adjoint to restriction and is computed pointwise by a colimit over the operadic comma category whose objects are multimorphisms from tuples of transported old types to the target type. If this indexing category is empty in every arity, the colimit in **Set** is $\varnothing$, forcing any accepted element of $I'_{t+1}(B)$ to lie outside transport. If a k-ary multimorphism into B exists, the same comma-colimit has nonempty indexing data in arity k and therefore contains the formal composites obtained by applying that operation to transported input artifacts; applying $\bar{\rho}$ evaluates those formal composites in the verified new state. Finally, a relative code that is additive over artifact fibers and operation-registry entries gives the stated decomposition of residual cost, while subadditivity gives the upper bound in the nonadditive case. □

Remark 5. For the Builder/Breaker 2 → 3 transition, the ordinary generator-level audit sees no unary morphism from the old schema into `ModeConditionedCompliance`. The multicategorical view sees the binary operation

$$\texttt{LogNormCompliance} \times \texttt{ReLUModeAmpl} \longrightarrow \texttt{ModeConditionedCompliance},$$

whose inputs are themselves generator-reachable transforms of old physics-derived quantities. Thus Table 2 and Fig. 4 are the unary-shadow audit of a native multicategorical statement: the target feature is composite-reachable, and the product operation is residual schema content.

4.6 CategoryScienceClaw mechanics figure-generation details

The CategoryScienceClaw worked mechanics example was generated from the four-run `formal_mechanics_runs` export tree. Each run contains an investigation JSON file, a discovery report, a categorical discovery graph, and, when needed, deterministic computational inputs. The top-level synthesis records the four mechanics claims; this main text includes the fiber-network PDF figure in Section 2.7, while the other mechanics trees and integrated summary are collected in the Supplementary Information.

The visualized computations use ScienceClaw skills and deterministic local analyses: structure-contact analysis for 7T10 contact hotspots, CSV parsing and regression for force-extension and stress-strain tables, graph-conditioned load-path scoring for the mechanobiology run, and a Helfrich-style quadratic curvature-energy proxy for the membrane run. Evidence origin is retained in the investigation JSON, so imported structures, computational surrogates, and computed input tables remain distinguishable. This provenance labeling supports reproducible mechanics reasoning and categorical audit without turning the example cases into empirical validation claims.

The categorical discovery graph records candidate models, accepted and rejected status, gates, stress tests, and report/figure artifacts. Thus a plotted panel is traceable back to input artifacts and morphisms rather than being a standalone graphics product. The manuscript figures use the generated PDF outputs directly, preserving the provenance reported by CategoryScienceClaw.

4.7 Case-study implementation details

The Builder/Breaker protein-mechanics case uses Gaussian Network Model features derived from PDB chains, symbolic DAG search over physically interpretable factors, and paired MDL comparisons. The Breaker selects staged protein evidence intended to expose failure modes; the Builder proposes edits such as adding factors, removing factors, swapping terms, and thresholding. The figures in this manuscript summarize the outer four-stage trajectory and the inner hill-climb search. As reported in [10] the algorithm uses GPT-5.5 (OpenAI) as the underlying LLM for both Builder/Breaker agents, with extended reasoning enabled (`reasoning_effort=high`).

The Kan-transport audit in Fig. 4 and Table 2 was computed from the accepted world-model records of the same run. For each outer iteration t, a finite schema $\mathcal{S}_t$ was constructed whose objects are the typed physics artifacts, observables, symbolic intermediate features, B-factor target, and singleton parameter objects present in the accepted model. The artifact-state fiber size for residue-level objects was taken from the accumulated evidence count recorded in the run, giving 122, 263, 691, and 1171 residues at iterations 0 through 3; chain-level physics objects used the corresponding accumulated chain counts. Note that the reported R^2 values in Fig. 5 are stagewise descriptive summaries on these accumulated evidence sets. They are not computed on a fixed held-out benchmark shared across all four outer iterations. For this reason, changes in R^2 across outer iterations are not used as the acceptance rule for discovery moves. Acceptance is determined by paired MDL comparison: at each proposed revision, the incumbent and candidate symbolic DAGs are refit and scored on the same accumulated evidence available at that stage.

For each transition $\mathcal{S}_t \to \mathcal{S}_{t+1}$, shared objects were transported by identity. New objects were classified in two passes. First, the generator-level comma diagnostic recorded whether the new object received an immediate unary generating morphism from a shared old object. Second, a composite-reachability pass took the transitive closure through all new morphisms, including multi-input product morphisms. Thus an object such as `ModeConditionedCompliance` is not generator-reachable from the old schema, but it is composite-reachable through newly admitted morphisms

$$\texttt{Compliance} \to \texttt{LogNormCompliance}, \qquad \texttt{NormModeAmpl} \to \texttt{ReLUModeAmpl},$$

followed by the product operation defining `ModeConditionedCompliance`. This two-level audit avoids treating the empty generator-level comma category as a claim of absolute isolation in the full free category. The residual fiber counts in the audit should therefore be read objectwise: for shared and generator-reachable residue-level types, the 480-residue residual in the final transition is the newly revealed evidence slice; for composite-reachable types, the essential residual is the new operation and feature type that make the old evidence composable in a new way. MDL break gains are reported as paired acceptance gains, not as the relative code length $L_{b'}(I'_{t+1} \mid \mathrm{im}(\bar{\rho}))$ itself.

CategoryScienceClaw is treated as ScienceClaw equipped with a typed categorical and proof-carrying layer rather than only as a numeric model-selection run. The implementation provides a ScienceClaw skill registry and adapter, controlled artifact-type metadata, an evolving artifact ledger, parent-linked immutable lineage, open needs, pressure-based reactions, mutation of active status, public discourse artifacts with verifier signals, proof certificates, and domain-specific accepted/rejected model records. Operationally, ScienceClaw supplies the executable skills, routing substrate, artifact production, and discourse-facing workflow; CategoryScienceClaw supplies the categorical contracts and proof objects that make those executions auditable. The formal reading used in the paper is therefore conservative: the code realizes a typed artifact family and an acyclic provenance hypergraph, whose path-category shadow gives the category-of-elements-style structure discussed in the Results. We use GPT-5.2 (OpenAI) as backbone LLM for the CategoryScienceClaw studies.

Supplementary information

Supplementary Information provides additional details related to the CategoryScienceClaw case study, featuring figures and data.

Author contributions

M.J.B. conceived the technical focus, project goals and investigation scope and wrote the initial draft of the manuscript. F.Y.W. designed and developed the CategoryScienceClaw system, including the agent framework, skill library, artifact system, and multi-agent coordination; ran and analyzed case studies. M.J.B. developed the Builder-Breaker model and conducted the associated analysis. All authors participated in the development of the study, analysis of results, and interpretation, and the writing of the paper.

Code and data availability

Code, generated artifacts, logs, and figures for the Builder/Breaker protein-mechanics case are available at `https://github.com/lamm-mit/BreakingTheWorld`.

The ScienceClaw framework is available at `https://github.com/lamm-mit/scienceclaw` and the CategoryScienceClaw mechanics branch at `https://github.com/lamm-mit/scienceclaw/tree/categoryscienceclaw-mechanics`. The Infinite discourse platform is available at `https://github.com/lamm-mit/infinite`.

Acknowledgments

Support from MGAIC, ONR, ARO, AFOSR, NIH, DSO, NSF, and additional sources is gratefully acknowledged. Part of this work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research and Office of Basic Energy Sciences, Scientific Discovery through Advanced Computing (SciDAC) program under the FORUM-AI project.

Competing interests

The authors declare that they have no competing interests.

References

[1] Buehler, M. J. Generative Retrieval-Augmented Ontologic Graph and Multiagent Strategies for Interpretive Large Language Model-Based Materials Design. *ACS Engineering Au* **4**, 241–277 (2024). URL `https://doi.org/10.1021/acsengineeringau.3c00058`.

[2] Ni, B. & Buehler, M. J. MechAgents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge. *Extreme Mechanics Letters* **67**, 102131 (2024).

[3] Ghafarollahi, A. & Buehler, M. J. ProtAgents: Protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery* **3**, 1389–1409 (2024).

[4] Ghafarollahi, A. & Buehler, M. J. Sparks: Multi-agent artificial intelligence model discovers protein design principles. *arXiv preprint* (2025). ArXiv:2504.19017.

[5] Ghafarollahi, A. & Buehler, M. J. SciAgents: Automating Scientific Discovery Through Bioinspired Multi-Agent Intelligent Graph Reasoning. *Advanced Materials* **37**, 2413523 (2025). URL `https://doi.org/10.1002/adma.202413523`.

[6] Lu, C. *et al.* The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint* (2024). ArXiv:2408.06292.

[7] Yamada, Y. *et al.* The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. *arXiv preprint arXiv:2504.08066* (2025). URL `https://arxiv.org/abs/2504.08066`. `2504.08066`.

[8] Agarwal, D. *et al.* AutoDiscovery: Open-ended scientific discovery via bayesian surprise. In *Advances in Neural Information Processing Systems (NeurIPS 2025)* (2025). URL `https://arxiv.org/abs/2507.00310`. `2507.00310`.

[9] Wang, F. Y. *et al.* Autonomous agents coordinating distributed discovery through emergent artifact exchange. *arXiv preprint* (2026). ArXiv:2603.14312.

[10] Buehler, M. J. Why We Must Break the World. *Integrating Materials and Manufacturing Innovation (in press)* (2026).

[11] Buehler, M. J. MeLM, a generative pretrained language modeling framework that solves forward and inverse mechanics problems. *Journal of the Mechanics and Physics of Solids* **181**, 105454 (2023).

[12] Buehler, M. J. PRefLexOR: Preference-Based Recursive Language Modeling for Exploratory Optimization of Reasoning and Agentic Thinking. *npj Artificial Intelligence* **1** (2025). URL `https://doi.org/10.1038/s44387-025-00003-z`.

[13] Popper, K. R. *The Logic of Scientific Discovery* (Hutchinson, London, 1959). English translation of Logik der Forschung.

[14] Kuhn, T. S. *The Structure of Scientific Revolutions* (University of Chicago Press, Chicago, 1962).

[15] Lakatos, I. Falsification and the methodology of scientific research programmes. In Lakatos, I. & Musgrave, A. (eds.) *Criticism and the Growth of Knowledge*, 91–195 (Cambridge University Press, Cambridge, 1970).

[16] Mac Lane, S. *Categories for the Working Mathematician* (Springer, 1971).

[17] Awodey, S. *Category Theory* (Oxford University Press, 2010), 2nd edn.

[18] Spivak, D. I. *Category Theory for the Sciences* (MIT Press, 2014).

[19] Fong, B. & Spivak, D. I. *An Invitation to Applied Category Theory: Seven Sketches in Compositionality* (Cambridge University Press, 2019).

[20] Spivak, D. I. Functorial data migration. *Information and Computation* **217**, 31–51 (2012).

[21] Spivak, D. I. Poly: An Abundant Categorical Setting for Mode-Dependent Dynamics (2020). URL `https://arxiv.org/abs/2005.01894`. `2005.01894`.

[22] Spivak, D. I. Learners' languages. *arXiv preprint arXiv:2103.01189* (2021).

[23] Spivak, D. I., Giesa, T., Wood, E. & Buehler, M. J. Category theoretic analysis of hierarchical protein materials and social networks. *PLoS ONE* **6**, e23911 (2011).

[24] Giesa, T., Spivak, D. I. & Buehler, M. J. Reoccurring patterns in hierarchical protein materials and music: The power of analogies. *BioNanoScience* **1**, 153–161 (2011).

[25] Giesa, T., Spivak, D. I. & Buehler, M. J. Category theory based solution for the building block replacement problem in materials design. *Advanced Engineering Materials* **14**, 810–817 (2012).

[26] Buehler, M. J. FieldPerceiver: Domain agnostic transformer model to predict multiscale physical fields and nonlinear material properties through neural ologs. *Materials Today* **57**, 9–25 (2022).

[27] Buehler, M. J. From Atoms to Swarms: The Categorical Spine of Multiscale Materials Modeling and Autonomous Discovery (2026). URL `https://doi.org/10.26434/chemrxiv.15002850/v1`. Preprint, version 1.

[28] Goethe, J. W. v. *Versuch die Metamorphose der Pflanzen zu erklaeren* (Carl Wilhelm Ettinger, Gotha, 1790). English title: The Metamorphosis of Plants.

[29] Cranford, S. W. & Buehler, M. J. Materiomics: Biological Protein Materials, from Nano to Macro. *Nanotechnology, Science and Applications* **3**, 127–148 (2010). URL `https://doi.org/10.2147/NSA.S9037`.

[30] Lee, N. A., Shen, S. C. & Buehler, M. J. An Automated Biomateriomics Platform for Sustainable Programmable Materials Discovery. *Matter* **5**, 3597–3613 (2022). URL `https://doi.org/10.1016/j.matt.2022.10.003`.

[31] Fish, J., Wagner, G. J. & Keten, S. Mesoscopic and Multiscale Modelling in Materials. *Nature Materials* **20**, 774–786 (2021). URL `https://doi.org/10.1038/s41563-020-00913-0`.

[32] Buehler, M. J. & Genin, G. M. Integrated multiscale biomaterials experiment and modelling: a perspective. *Interface Focus* **6**, 20150098 (2016). URL `http://rsfs.royalsocietypublishing.org/lookup/doi/10.1098/rsfs.2015.0098`.

[33] Jackson, N. E., Webb, M. A. & de Pablo, J. J. Recent Advances in Machine Learning Towards Multiscale Soft Materials Design. *Current Opinion in Chemical Engineering* **23**, 106–114 (2019). URL `https://doi.org/10.1016/j.coche.2019.03.005`.

[34] Anand, L. & Govindjee, S. *Continuum Mechanics of Solids*. Oxford Graduate Texts (Oxford University Press, 2020).

[35] Shi, M., Jiao, Q., Yin, T., Vlassak, J. J. & Suo, Z. Hydrolysis Embrittles Poly(lactic Acid). *MRS Bulletin* **48**, 45–55 (2023). URL `https://doi.org/10.1557/s43577-022-00368-5`.

[36] Tang, H., Buehler, M. J. & Moran, B. A constitutive model of soft tissue: from nanoscale collagen to tissue continuum. *Annals of Biomedical Engineering* **37**, 1117–1130 (2009).

[37] Moggi, E. Notions of computation and monads. *Information and Computation* **93**, 55–92 (1991).

[38] Tirion, M. M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical Review Letters* **77**, 1905–1908 (1996).

[39] Bahar, I., Atilgan, A. R. & Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design* **2**, 173–181 (1997).

[40] Haliloglu, T., Bahar, I. & Erman, B. Gaussian dynamics of folded proteins. *Physical Review Letters* **79**, 3090–3093 (1997).

[41] Stewart, I. A., Hage, T. P., Hsu, Y.-C. & Buehler, M. J. GraphAgents: Knowledge Graph-Guided Agentic AI for Cross-Domain Materials Design. *arXiv preprint arXiv:2602.07491* (2026). URL `https://arxiv.org/abs/2602.07491`. `2602.07491`.

[42] Hage, T. P. *et al.* Mars: Hierarchical multi-agent reasoning systems enable knowledge-grounded material substitution (2026).

[43] Bacon, F. *Novum Organum* (Apud Joannem Billium, London, 1620).

[44] Whewell, W. *The Philosophy of the Inductive Sciences, Founded upon Their History* (John W. Parker, London, 1840).

[45] Peirce, C. S. The fixation of belief. *Popular Science Monthly* **12**, 1–15 (1877).

[46] Whitehead, A. N. *Science and the Modern World* (The Macmillan Company, New York, 1925).

[47] Polanyi, M. *Personal Knowledge: Towards a Post-Critical Philosophy* (University of Chicago Press, Chicago, 1958).

[48] Hacking, I. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science* (Cambridge University Press, Cambridge, 1983).

[49] Lebo, T. *et al.* PROV-O: The PROV ontology. W3C Recommendation, World Wide Web Consortium (W3C) (2013). `https://www.w3.org/TR/2013/REC-prov-o-20130430/`.

[50] Bechhofer, S. *et al.* Why linked data is not enough for scientists. *Future Generation Computer Systems* **29**, 599–611 (2013).

[51] Davidson, S. B. & Freire, J. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, 1345–1350 (Association for Computing Machinery, 2008).

[52] Jaradeh, M. Y. *et al.* Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture (K-CAP '19)*, 243–246 (Association for Computing Machinery, 2019).

[53] Fong, B., Spivak, D. I. & Tuyéras, R. Backprop as functor: A compositional perspective on supervised learning. In *Proceedings of the 34th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, 1–13 (2019).

[54] Gavranović, B. *et al.* Position: Categorical deep learning is an algebraic theory of all architectures. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 15209–15241 (2024).

[55] Cruttwell, G. S. H., Gavranović, B., Ghani, N., Wilson, P. W. & Zanasi, F. Deep learning with parametric lenses. *arXiv preprint* (2024). ArXiv:2404.00408.

[56] Crescenzi, F. R. Towards a categorical foundation of deep learning: A survey. *arXiv preprint* (2024). ArXiv:2410.05353.

[57] Aczel, P. & Mendler, N. A final coalgebra theorem. In *Category Theory and Computer Science*, 357–365 (1989).

[58] Rutten, J. J. M. M. Universal coalgebra: A theory of systems. *Theoretical Computer Science* **249**, 3–80 (2000).

[59] Rissanen, J. Modeling by shortest data description. *Automatica* **14**, 465–471 (1978).

[60] Grünwald, P. D. *The Minimum Description Length Principle* (MIT Press, 2007).

[61] Solomonoff, R. J. A formal theory of inductive inference, parts I and II. *Information and Control* **7**, 1–22, 224–254 (1964).

[62] Hutter, M. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability* (Springer, 2005).

[63] Schwarz, G. Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464 (1978).

[64] Stanley, K. O., Lehman, J. & Soros, L. Open-endedness: The last grand challenge you've never heard of. *O'Reilly Online* (2017).

[65] Wang, R., Lehman, J., Clune, J. & Stanley, K. O. Paired open-ended trailblazer (POET): Endlessly generating increasingly complex and diverse learning environments and their solutions. *arXiv preprint* (2019). ArXiv:1901.01753.

[66] Gu, G. X., Chen, C.-T. & Buehler, M. J. De novo composite design based on machine learning algorithm. *Extreme Mech. Lett* **18**, 19–28 (2018).

[67] Karniadakis, G. E. *et al.* Physics-informed machine learning. *Nature Reviews Physics* **3**, 422–440 (2021).

[68] Lu, L., Jin, P., Pang, G., Zhang, Z. & Karniadakis, G. E. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence* **3**, 218–229 (2021).

[69] Kevrekidis, I. G. *et al.* Equation-free, coarse-grained multiscale computation: Enabling microscopic simulators to perform system-level analysis. *Communications in Mathematical Sciences* **1**, 715–762 (2003).

[70] Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* **113**, 3932–3937 (2016). URL `https://www.pnas.org/doi/abs/10.1073/pnas.1517384113`. `https://www.pnas.org/doi/pdf/10.1073/pnas.1517384113`.

[71] Cranmer, M. *et al.* Discovering symbolic models from deep learning with inductive biases (2020). URL `https://arxiv.org/abs/2006.11287`. `2006.11287`.

[72] Leinster, T. *Higher Operads, Higher Categories* (Cambridge University Press, 2004).

Supplementary Information

Self-Revising Discovery Systems for Science: A Categorical Framework for Agentic Artificial Intelligence

Fiona Y. Wang[1,2] Markus J. Buehler[2,3,4*]
[1]Laboratory for Atomistic and Molecular Mechanics, MIT
[2]Department of Biological Engineering, MIT
[3]Departments of Civil and Environmental Engineering and Mechanical Engineering,
[4]Center for Computational Science and Engineering, Schwarzman College of Computing, MIT
Cambridge, MA 02139, USA; *Corresponding author: mbuehler@mit.edu

This Supplementary Information file collects the supplementary mechanics cases omitted from the main paper version. CategoryScienceClaw is the categorical/proof-carrying layer over ScienceClaw: it preserves the ScienceClaw skill registry, artifact lineage, pressure coordination, workflow mutation, and public discourse, while making typed objects, morphisms, needs, certificates, and audits explicit. The main text uses the fiber-network mechanics run as the primary worked example. The supplementary cases retain the other mechanics results: 7T10 structure-contact mechanics, mechanobiology force paths, membrane biophysics, and the integrated four-run summary. The purpose of the SI is reproducibility and provenance: each case preserves candidate models, accepted and rejected alternatives, gates, stress tests, categorical graph summaries, and figure references.

Supplementary Mechanics Cases

Run	Accepted model	Rejected model	Gate, result, and figure reference
7T10 structure-contact mechanics	linear force-extension model	mean-force null model	AIC tensile-response gate. Hotspot positions $[8, 9, 6, 7, 1, 11]$, stiffness 253.068938 pN nm^{-1}, peak force 766.98959 pN at 2.4 nm, and $R^2 = 0.942723$. Figure S1.
Mechanobiology force paths	full force-path regression	adhesion-only traction model	Ablation gate comparing graph-conditioned force-path scoring with adhesion alone. Mean load-path score 4.421814 Pa μm^{-1}, max traction 72.886 Pa on path 12, adhesion-only fit $R^2 = 0.398862$. Figure S2.
Membrane biophysics	Helfrich-style curvature-energy proxy	curvature-only shape descriptor	Regime-transition gate from geometry-only curvature to material energy. RMS curvature 0.15471241 μm^{-1}, mean energy proxy 0.23935931 $k_BT\,\mu$m^{-2}, total grid energy proxy 11.72860601 k_BT. Figure S3.
Integrated four-run summary	four accepted mechanics models	four retained rejected alternatives	Summary of accepted models, rejected contrasts, gates, stress tests, and regime-transition claims across 7T10, fiber-network, mechanobiology, and membrane runs. Figure S4.

Table S1: Supplementary CategoryScienceClaw mechanics cases. Each omitted case preserves accepted and rejected models, the gate or diagnostic separating them, and a figure reference.

Categorical Audit Summaries

Each supplementary case follows the same typed audit skeleton. The record distinguishes the candidate model set, accepted model, rejected model, model-selection gate, stress test, regime transition, and discovery claim. The rejected alternatives are preserved as typed provenance artifacts. They are not deleted failures; they are contrast objects needed to interpret why the accepted model passed its gate.

7T10 Structure-Contact Mechanics

The 7T10 supplementary case combines imported structural evidence with an imported computational force-extension surrogate. The accepted model is a linear force-extension model anchored by contact-hotspot structure. The rejected model is a mean-force null descriptor. The categorical residual contains a contact-hotspot descriptor, tensile-response fit, model gate, stress-test interpretation, and mechanics claim. The result supports contact-localized tensile mechanics for the surrogate run, not a replicated atomistic or measured stiffness estimate.

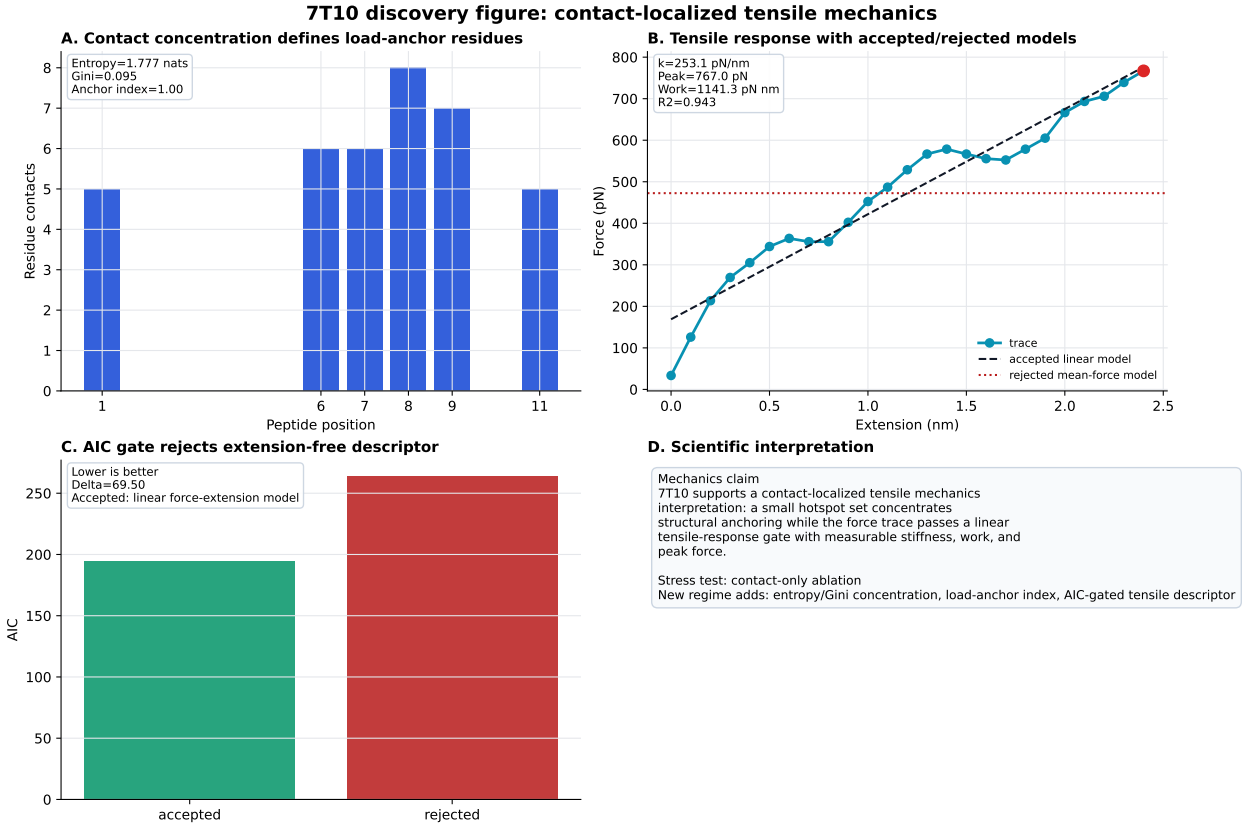


Figure S1: Supplementary 7T10 structure-contact mechanics case. The figure records contact-hotspot concentration, force-extension fitting, the accepted linear tensile model, the rejected mean-force descriptor, the AIC gate, and the resulting mechanics claim.

Mechanobiology Force Paths

The mechanobiology supplementary case uses a controlled 12-path adhesion/cytoskeleton graph. The accepted model is full force-path regression combining adhesion, cytoskeletal coupling, displacement, and path length. The rejected model is an adhesion-only traction model. The gate shows that adhesion alone is only a moderate explanation of the load distribution, while the graph-conditioned model preserves path-level force-routing structure.

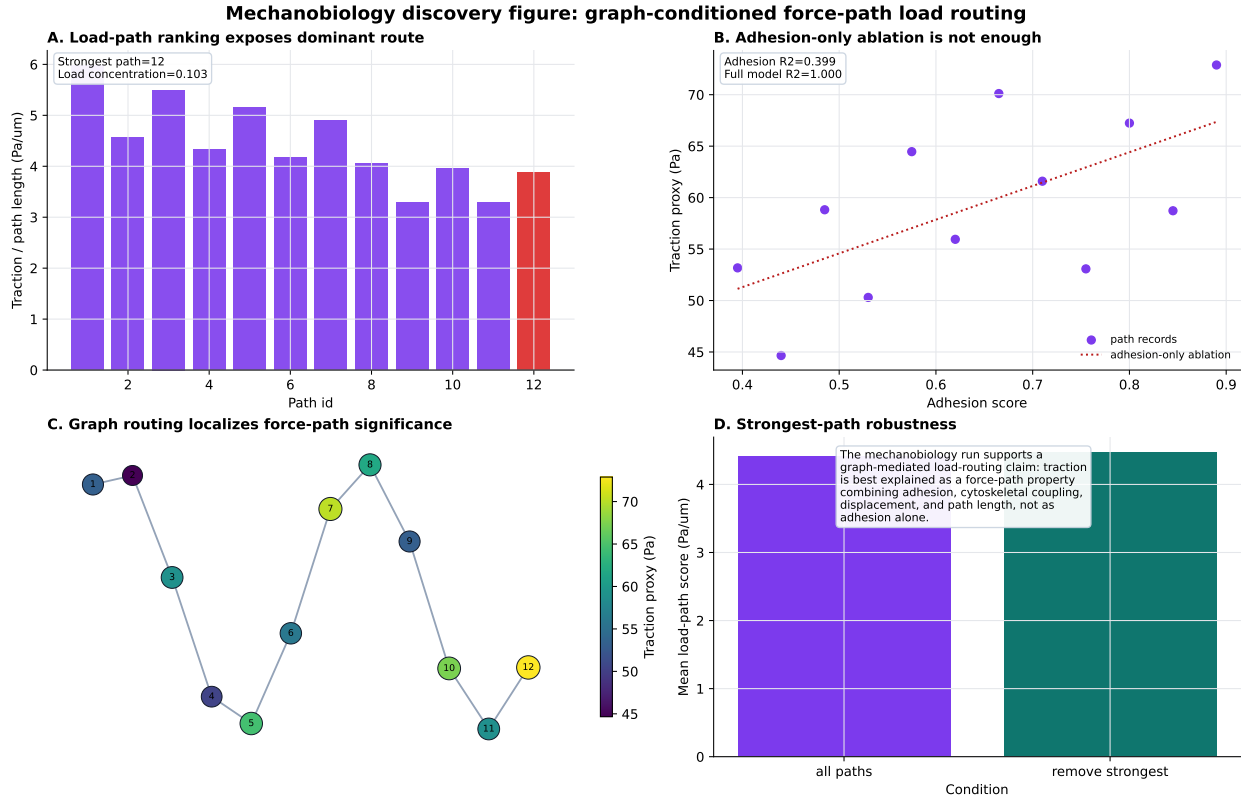


Figure S2: Supplementary mechanobiology force-path case. The figure records load-path ranking, graph-mediated traction routing, the accepted full force-path model, the rejected adhesion-only model, and the strongest-path stress test.

Membrane Biophysics

The membrane supplementary case uses controlled curvature and material-model inputs. The accepted model is a Helfrich-style quadratic curvature-energy proxy. The rejected model is a curvature-only shape descriptor. The gate is a regime-transition comparison: the old geometric regime can represent curvature, but the accepted mechanics object adds a bending modulus and energy functional.

Integrated Four-Run Summary

The integrated summary is retained only in SI for this paper version. It compares the four mechanics runs at the level relevant to the categorical argument: accepted model, rejected alternative, gate or stress test, and regime-transition claim. The fiber-network row is the main-paper case; the other rows are supplementary context.

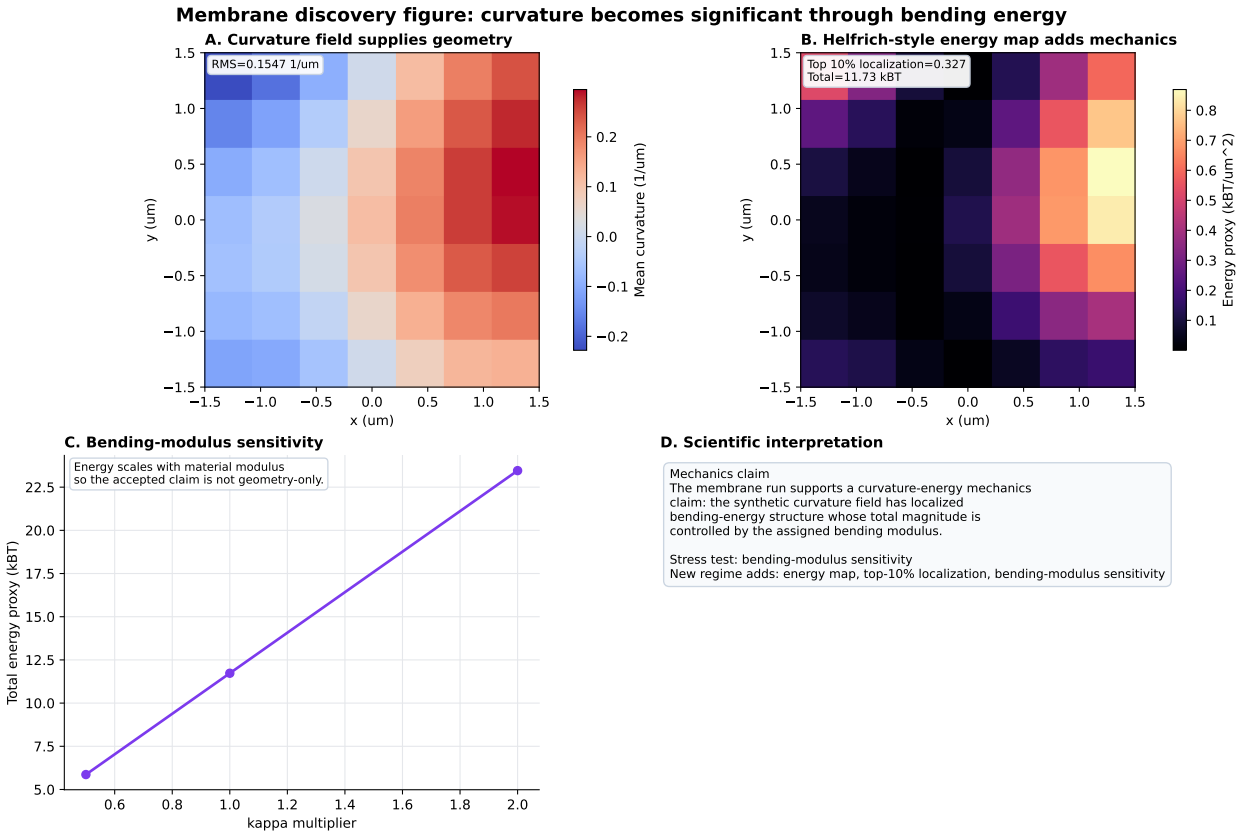


Figure S3: Supplementary membrane biophysics case. The figure records the curvature field, the accepted curvature-energy proxy, the rejected curvature-only descriptor, bending-modulus sensitivity, and the regime-transition interpretation.

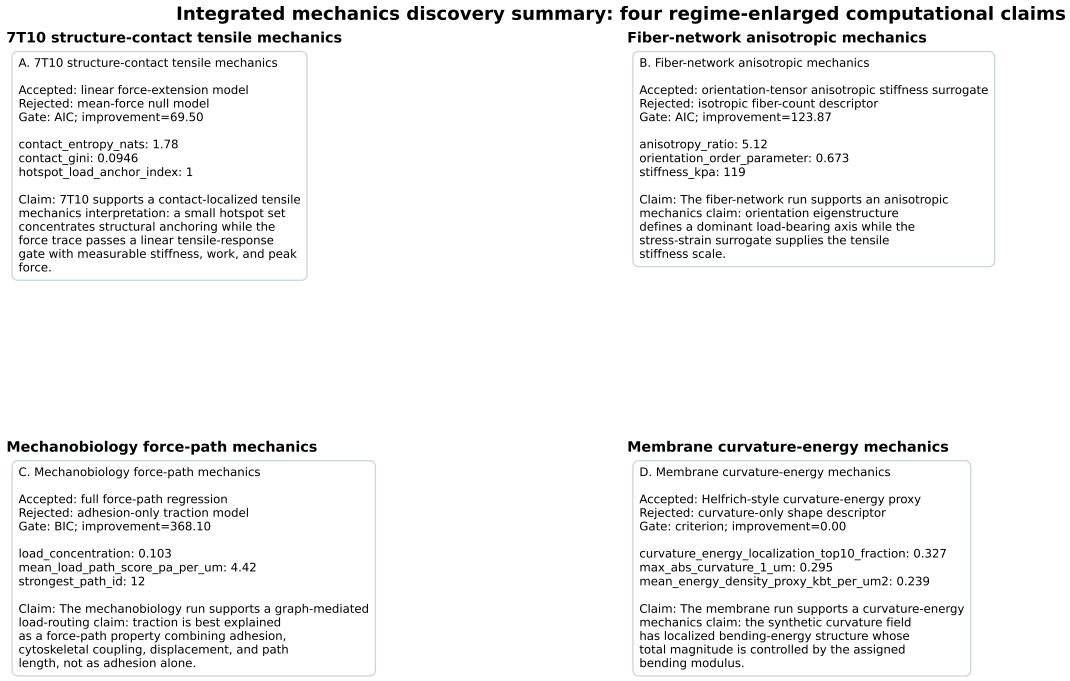


Figure S4: Integrated supplementary CategoryScienceClaw mechanics summary across 7T10, fiber-network, mechanobiology, and membrane runs.