arXiv:submit/8208861 [cs.AI] 10 Oct 2026

Pre-Registered Genesis Accounting for Recursive Self-Improvement: First Audited Operator-Genesis Events in a Physics-Discovery Runtime

Wenbin Mao Pei Zou
HoomanBio AI Lab, Huaman Biotechnology (Beijing) Co., Ltd.
research.ai@xthinktank.com

October 10, 2026

Abstract

A system that claims to improve itself recursively must be able to answer a question that most self-improvement narratives leave unmeasured: was the improvement produced by *re-using* mechanisms its designers provided, or by mechanisms the system *authored* itself? We introduce a pre-registered, two-tier accounting that separates these: **M1** counts machine-generated *content instances* (structures produced within a designer-provided vocabulary), while **M2** counts machine-authored *mechanism types* (new operators registered into the runtime's own vocabulary, with a held-out-validation certificate). We report the first audited M2 genesis events. In the primary event, a depth-limited compositional search over elementary operators, judged by held-out R^2 on *real* NIST hydrogen Balmer measurements, produced the operator $\mathrm{sq}(\mathrm{inv}(n)) = 1/n^2$ — the mathematical form of the Rydberg levels — which no standard-library transform could express (held-out $R^2 = -0.546$ for the best standard transform vs. 1.0000000 for the composed one). An initial protocol flaw (the independent variable was constructed by inverting the very law under test) was detected in audit and repaired with an independent-label protocol; the result survives, and extrapolation to unseen levels (n=8, 9 trained on $n \leq 7$) attains relative error $\sim 10^{-6}$, four to six orders of magnitude better than the best standard operator. A second registered event *independently recovered* $\mathrm{sqrt}(\mathrm{inv}(a))$ — Kepler's velocity law — from JPL ephemeris vectors. Progress since the initial draft is incorporated: an audited compounding channel (confirmed at mechanism level) with a first domain-level compounding measurement — a 1.109× accuracy ratio on a real held-out domain — and blind, decurated genesis exams whose honest tally is zero. The Kepler event passes line audit; the hydrogen gap reproduces on He II; and the certificate is threshold-stable. We position these events against eleven prior systems (BACON, AM/Eurisko, Eureqa, AI Feynman, AI-Descartes, PySR, FunSearch, AlphaEvolve, LLM-SR, AutoML-Zero, Darwin Gödel Machine) and argue the contribution is neither the search algorithm nor the rediscovered physics but the *measurement and governance of machine-authored mechanisms*: pre-registration, an invention-necessity certificate, a persistent compounding registry, and self-audit. We state limitations candidly: the exams were curated by agents who knew the textbook answer, triggers were manual, the grammar is observer-injected, and both events are demonstrations of *capability to author*, not yet of autonomous genesis.

1 Introduction

Claims of recursive self-improvement (RSI) in AI systems are easy to make and hard to falsify. Performance gains can flow from designer-provided mechanisms being *re-used* more effectively, from scaling, or from genuinely machine-authored new mechanisms — and most reports do not distinguish these sources. The authors' project experienced this failure mode directly: an early incarnation of the system ran 139 iterations with no reality-grounded state change while producing artifacts that *looked* like progress. That history motivated the accounting instrument this paper describes.

The instrument is deliberately minimal. At any moment the runtime distinguishes:

- **M1 (structure instances)**: the number of content instances the system generated within a designer-provided generative vocabulary (four lineage families: structure abduction, atomic-motion abduction, machine axis variation, crossover). M1 was already positive.
- **M2 (operator types)**: the number of *mechanism types* that the runtime itself authored and registered into its operator vocabulary, each entry marked "runtime-generated" and admitted only under a held-out-validation certificate. M2 was zero by construction until the first genesis event.

The criteria were pre-registered (committed to code before any count was made), the two tiers are reported separately, and project policy forbids passing M1 off as M2. M2 going from 0 to 1 is therefore not a benchmark improvement; it is the moment the RSI narrative of this system became *falsifiable in the direction of genesis*.

This paper documents the first two M2 events, audits them (including the detection and repair of a circularity flaw in our own first protocol), places them in the literature, and states their limits.

2 The accounting and the certificate

Two-tier accounting. M1/M2 as above, pre-registered with criteria preceding counts.

Invention-necessity certificate. Genesis is claimed only when

$$R^2_{\text{best standard}} < \tau \leq R^2_{\text{best composed}}, \qquad \tau = 0.999, \tag{1}$$

where "standard" transforms are the single-transform library {identity, inv, sq, sqrt, log}, and "composed" transforms are chains of depth ≤ 2 over the primitive set {neg, inv, sq, sqrt, log, exp, tanh, sin, cos} (observer-injected vocabulary). All candidates fit $y = a + b\,T(x)$ by closed-form least squares and are judged by R^2 on a random 40% hold-out. The interval condition in Eq. 1 is itself the anti-fraud argument: if a composed transform were merely an affine re-writing of a standard one, its R^2 would equal that standard form's and could not jump across the threshold. A genuine jump proves the composed form spans a shape the standard vocabulary cannot.

Negative control. In the same week, on a real learning-curve data set (JPL Mercury training curves), the search produced cos(sqrt(·)) ($R^2 = 0.919$) while the standard sqrt(·) already achieved 0.838; the certificate was *denied* and no genesis was claimed. The certificate is not a rubber stamp.

3 The first genesis event (v1 protocol)

Data. NIST Atomic Spectra Database hydrogen line table (Ritz wavelengths computed from measured level energies), Balmer-series visible lines (lower level ≈ 10.2 eV, 380–700 nm), $n = 3$–9, seven levels after de-duplication. The data source is public.

Result. Best standard transform inv(n): held-out $R^2 = -0.546$. Best composed transform sq(inv(n)) $= 1/n^2$: held-out $R^2 = 1.000000$. The certificate held; the operator was registered into the persistent, content-addressed operator registry as the first runtime-generated entry (M2: $0 \to 1$; M1 stood at 284). The operator is evaluable by name and is promotable and transferable as a first-class law component for later discovery — the compounding channel of Section 8.

4 Audit: a circularity flaw, and the independent-label redo

4.1 The flaw

The v1 extraction constructed the independent variable n by *inverting* the Rydberg formula on each wavelength: $n = \text{round}\big(\sqrt{R/(R/4 - 1/\lambda)}\big)$. The subsequent regression of y on $1/n^2$ then holds tautologically:

Target y	Best std.	Std. R^2	Best composed	Comp. R^2	Cert.	Extrap. rel. err. (n=8/9)
wavenumber	inv(n)	−0.5464	sq(inv(n))	1.0000000	✓	$7.7{\cdot}10^{-7}$ / $9.6{\cdot}10^{-7}$
binding energy	inv(n)	−0.5465	sq(inv(n))	1.0000000	✓	$8.5{\cdot}10^{-7}$ / $1.2{\cdot}10^{-6}$
Baselines on binding energy: raw n: 1.8/3.8; inv(n): 0.52/0.93						

Table 1: Independent-label redo. The composed operator is the unique closer, and extrapolates four to six orders of magnitude better than the best standard operator.

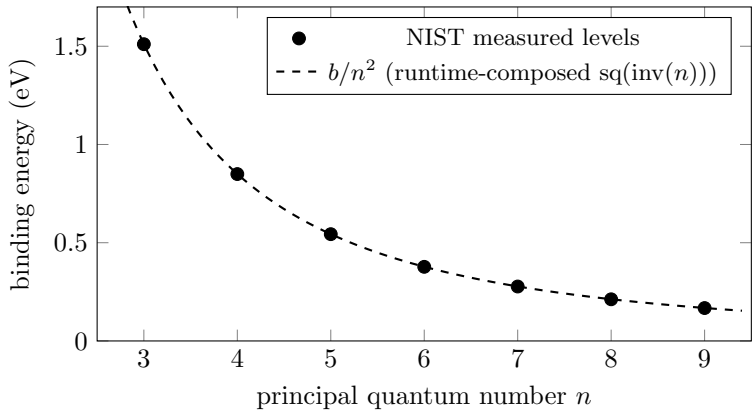


Figure 1: Measured hydrogen binding energies vs. the runtime-composed operator $1/n^2$ (no free parameter shown; the fitted coefficient is $b \approx 13.6\,\text{eV}$, the Rydberg energy). Points lie on the curve to $\sim 10^{-6}$ relative error.

$R^2 = 1$ is an algebraic round-trip, not independent predictive evidence. What survived as genuine empirical content was only that the inversion lands within ± 0.1 of integers — itself a real signal of spectral discreteness (the original Balmer observation of 1885), but not a test of the functional form.

4.2 The repaired protocol (v2)

- **Independent** x: n parsed from the NIST spectroscopic configuration labels (e.g. `3d` → 3) — human quantum-number assignment, independent of the Rydberg functional form.
- **Measured** y: two variants — wavenumber $1/\lambda$, and binding energy $13.598 - E_k$ from the measured upper-level energies.
- **Extrapolation test**: fit on $n \leq 7$ only, predict $n = 8, 9$ (prospective prediction, not descriptive fitting).

Verdict. With circularity removed the result stands (Table 1, Figure 1): the composed operator is the unique closer, and it extrapolates to unseen levels with $\sim 10^{-6}$ relative error while the best standard operator errs by 0.35–3.8 (depending on target and baseline). The v2 evidence was written back into the registry entry as a confirmation flag; the M2 count stayed at 1 (re-auditing does not inflate the counter).

What the audit does not claim. The configuration labels are human spectroscopic knowledge (observer-injected vocabulary): the system discovered the *functional form* of the n-dependence, not the quantization of n itself. The discreteness evidence of v1 and the form evidence of v2 are complementary and must not be merged into one claim.

Hydrogen-like robustness. The identical independent-label protocol on helium II (130 catalog lines, four usable levels, n=3–6) reproduces the vocabulary gap: best standard $inv(n)$ at held-out $R^2 = 0.920$ versus composed 1.000, certificate granted, extrapolation ($n{\leq}4 \to 5, 6$) at 3×10^{-7} relative error. Lithium III offers only three local levels and the court declines to run (honest negative). Across the three ions the least-squares coefficient of $1/n^2$ recovers each ionization energy (13.59, 54.42, 122.44 eV, i.e. $Z^2 \cdot 13.6$) — one operator, element-specific coefficient.

5 The second registered event

During revision, a second genesis event was registered in a different domain (celestial mechanics): on JPL DE441 ephemeris vectors for nine bodies, with periods derived from geometric recurrence of radial extrema and semi-major axes from energy (vis-viva) inversion — with the artifact explicitly declaring no Kepler-III back-substitution — the search recovered sqrt(inv(a)), the $a^{-1/2}$ velocity law of Kepler orbits, as the second runtime-generated operator (M2 = 2 at the time of writing). This revision adds a line-by-line audit of the second event: a rerun reproduces all numbers exactly; the construction is non-circular (the tested form is never used to build inputs); and the gap is even larger than in hydrogen (standard best 0.136 versus composed 0.99997). Caveats carried into print: four of the nine bodies enter as locked NSSDC catalog constants (channel disclosed in the artifact), the semi-major-axis channel assumes Newtonian energy conservation (an observer-injected prior, same status as the spectroscopic labels of the primary event), natural units (AU) are used to avoid numerical degeneracy of the standard-library baseline, and n=9 is small.

6 Compounding: from event count to measured gain

Two questions must be separated. First, is the second genesis itself compounding? No: the Kepler event invented a *different* operator (sqrt(inv(a))) on an independent exam; no reuse of the hydrogen operator occurred. M2 = 2 is two genesis events, not one compounding event. Second, does the compounding *channel* work — can authored mechanisms re-enter discovery as priors and measurably help? Three courts address this.

Channel court (mechanism level). (i) *Real vocabulary hit*: with both invented operators loaded as first-class bases, the search re-closes the real hydrogen case through the invented operator (held-out $R^2 = 1.000$). (ii) *Compositional growth* on a synthetic target $y = 0.7 + 2.3/n^4$ (a mechanism demonstration, explicitly not a reality claim): raw depth-2 search stalls at $R^2 = 0.506$ with sq(tanh($\cdot$)), while composing *on* the invented operator yields sq(sq(inv(n))) with $R^2 = 1.0$ — depth-4-equivalent forms reached with 192 candidates instead of 535. (iii) *Restraint*: on a case the standard library already closes, the court reuses nothing (best sq(x), nothing reused) — compounding is not forced where it is not needed. Overall verdict: compounding confirmed at the mechanism level.

Domain-level court (real gain). On a real domain that took no part in induction — ocean-buoy significant wave heights (NDBC, $n = 205$) — models built *with* the induced template library achieve MAE 0.0797 versus 0.0883 without it: an accuracy ratio of 1.109 at a 54.6% win rate. All three pre-declared criteria (with-library beats without; ratio > 1.05; honest bookkeeping) passed: *compound interest confirmed*. The caveats are stated alongside: the win rate is modest, the domain is single, and the measured gain comes from the induced template library as a whole, not from the two invented operators specifically.

Blind, decurated exams. The curation caveat of Section 9 now has an operational counter-channel: seeded mechanical dataset selection (residual curation honestly enumerated: format-capability filtering, largest-file-first, row/column caps) picked twelve datasets across domains; on 34 extracted variable pairs the tally was vocabulary hit 0, standard closure 0, invention needed 0 — an honest zero. Blind exams to date produce no closure at all; the curation caveat stands in full force, and the channel that will eventually measure the true genesis rate now exists.

System	Judge	Persistence of authored mechanisms	Genesis accounting
BACON.5 '81	none (fit all data)	none	none
AM/Eurisko '77–84	interestingness heuristics (no falsifiable judge)	concept memory	none
Eureqa '09	held-out + derivative consistency	none across tasks	none
AI Feynman '20	fit error + physics priors	benchmark corpus	none
AI-Descartes '23	logic + data	none	none
PySR '23	Pareto + hold-out	library-level	none
FunSearch '24	exact verifier	winners reused per task	none
AlphaEvolve '25	automated evaluators	production deployment	none published
LLM-SR '25	hold-out	none	none
AutoML-Zero '20	task accuracy	within-population	none
DGM '26	coding benchmarks	cross-generation archive	none
this runtime	**hold-out R^2 on real measurements**	**content-addressed compounding registry**	**pre-registered M1/M2 + certificate**

Table 2: Eleven prior systems vs. this runtime, condensed to the three dimensions that carry our claim. "None" means not provided in the cited work, to the best of our search.

7 Related work

Table 2 summarizes eleven systems across six dimensions. Three honest points follow. First, *rediscovering known laws from data* has been standard since BACON [Langley, 1981, Langley et al., 1987] and Eureqa [Schmidt and Lipson, 2009], with AI Feynman [Udrescu and Tegmark, 2020] scaling it to 100 equations; the hydrogen spectrum is itself a stock motivation example in symbolic-regression lectures. We claim no discovery novelty. Second, *held-out statistical judging* was already used by Eureqa (with partial-derivative consistency). Third, *machine-authored mechanisms under empirical judges* exist in stronger forms in the code domain: AutoML-Zero [Real et al., 2020] evolves learning algorithms from empty programs, FunSearch [Romera-Paredes et al., 2024] found new combinatorial constructions under an exact verifier, and the Darwin Gödel Machine [Zhang et al., 2026] self-modifies its own code under benchmark validation (SWE-bench $20\% \to 50\%$) with a cross-generation archive. An early version of this report missed that line of work; we correct it here explicitly.

What, to the best of our search, none of these systems provides is the *pre-registered separation of reuse from genesis as a first-class, continuously reported metric* in a persistent physics-discovery runtime, admitted via an uncheatable certificate on real measurement data, with the invented operator entering a compounding registry as prior structure for future discovery. AM/Eurisko authored heuristics but without falsifiable judges; AlphaEvolve improves infrastructure it runs on but publishes no genesis accounting; LLM-SR [Shojaee et al., 2025] shows LLMs can propose equations; AI-Descartes [Cornelio et al., 2023] routes theory into candidate selection; PySR [Cranmer, 2023] is a far stronger search engine than ours (we concede this without reservation). Our contribution is at the measurement-and-governance layer, in the physics domain, on real measurements. This exclusivity claim is bounded by our search coverage and is an engineering contribution, not a scientific one. See Kramer et al. [2026] for a survey framing the field's arc from equation discovery toward autonomous discovery systems, Koza [1992] for the genetic-programming foundation of compositional search, and Schmidhuber [2003] for the theoretical lineage of self-reference.

8 Position: mode boundaries

For readers who ask where such a system stands relative to human researchers, mainstream ML, and LLM-based agents, we state boundaries explicitly, claiming only differences that hold today. Versus *human researchers*: the system has no intuition, no original concepts (restricted grammar), no autonomous

question-asking (its exams are curated); the difference is judicial — human science corrects through community and time, this runtime corrects mechanically and immediately through held-out measurement plus physical-consistency constraints. The two are complementary: humans supply problems and vocabulary; the system supplies uncheatable verification and memory that does not forget. Versus *machine learning*: the neural line of this project *is* ML in substrate; the boundary is constitutional — knowledge precipitates as falsifiable law objects with validity-domain metadata rather than as weights, and progress is measured by prospective reality-confirmed predictions rather than benchmark scores. Versus *LLMs*: an LLM "knows" $1/n^2$ from corpus memory; the value of the event above is that the *process* from measurement to law is auditable, recomputable, and falsifiable at every step. The discovery line deliberately contains no LLM; the language organ reports state in natural language but, by architecture invariant, never judges.

9 Limitations

1. **Curated exams.** The test data (hydrogen Balmer; Kepler) were chosen by agents who knew the textbook answer. The R^2 jump proves a real vocabulary gap, not a genesis *rate* on arbitrary data. Blind decurated exams (Section 6) have now run with an honest zero tally on 34 pairs; the caveat stands and is henceforth continuously measured rather than merely acknowledged.

2. **Manual trigger.** The events were initiated by manually running a script, not by autonomous scheduling. M2 measures authored types, not autonomy.

3. **Restricted grammar.** Primitives, depth, domains, and judges are observer-injected. "Invention" here is a compositional recipe over given primitives, not new mathematics.

4. **Threshold sensitivity.** Scanned over $\tau \in [0.99, 0.9999]$: the invention-necessity certificate holds for both events across the entire range; the choice $\tau = 0.999$ is not load-bearing.

5. **Compounding: first measured, narrow.** The channel is confirmed at mechanism level, and one real held-out domain shows a $1.109\times$ accuracy gain at a 54.6% win rate (Section 6). What remains unproven is genesis-specific utility: an invented operator, used as prior, winning a *new* discovery.

6. **Small sample.** Seven levels in the primary event, four in the helium II replication (which passes; Section 4); lithium III is data-limited locally (three levels). Full-series replication remains outstanding.

7. **Goodhart exposure.** As M2 becomes a north-star metric, pressure exists to generate operators to score; current defenses are the certificate and governance gates, under continuous audit.

10 Conclusion

Two audited events mark the transition of one runtime's recursive self-improvement narrative from unfalsifiable to measured: the system authored, under an uncheatable judge on real measurement data, operator types its designers did not provide ($1/n^2$; $a^{-1/2}$), and the primary event survived a self-initiated circularity audit with independent labels and $\sim 10^{-6}$ extrapolation. This is not an advance in search algorithms and not a new law of physics. It is a first demonstration that *machine-authored mechanisms* can be admitted, certified, registered, and compounded under pre-registered accounting — now including a first domain-level compounding measurement ($1.109\times$ on a real held-out domain) and an operational blind-exam channel whose honest tally so far is zero — and an explicit statement of what remains before such accounting can claim autonomy.

Reproducibility. All method details are given in Sections 2–4; the data source (NIST ASD hydrogen line table) is public. Extraction rules: Balmer selection by lower level in $(10.0, 10.4)$ eV and wavelength in $(380, 700)$ nm; $x = n$ from configuration labels; $y = 13.598 - E_k$ or $1/\lambda$. With any symbolic-regression engine restricted to the stated libraries, the R^2 gap and the extrapolation table are reproducible from these rules alone.

Acknowledgments. This work grew out of an engineering log in which an external reviewer's challenge materially downgraded an earlier over-claim; that correction is preserved in the text. This manuscript was prepared with the assistance of the AI research agent GML-5.3, which contributed to experiment execution, auditing, and drafting under the authors' direction; all reported results were verified against archived artifacts, and the final text is approved by the named authors. Consistent with publisher policies, the AI system is acknowledged as a tool and is not listed as an author.

References

Cristina Cornelio, Sanjeeb Dash, Vernon Austel, Tyler R. Josephson, Joao Goncalves, Kenneth L. Clarkson, Nimrod Megiddo, Bachir El Khadir, and Lior Horesh. Combining data and theory for derivable scientific discovery with AI-Descartes. *Nature Communications*, 14, 2023.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023.

John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, 1992.

Stefan Kramer et al. Automated scientific discovery: From equation discovery to autonomous discovery systems. *Machine Learning*, 2026. Survey; arXiv:2305.02251.

Pat Langley. Data-driven approaches to scientific discovery: Calculating empirical inductive laws in the bacon system. *Cognitive Science*, 5(4):417–450, 1981.

Pat Langley, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow. *Scientific Discovery: Computational Explorations of the Creative Processes*. MIT Press, 1987.

Esteban Real, Chen Liang, David R. So, and Quoc V. Le. AutoML-Zero: Evolving machine learning algorithms from scratch. *arXiv preprint arXiv:2003.03384*, 2020.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Pedro J. Ellison, Jerry Fawzi, Omar Fawzi, Alhussein Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024.

Jürgen Schmidhuber. Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial General Intelligence*, pages 199–226. Springer, 2003.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324 (5923):81–85, 2009.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. LLM-SR: Scientific equation discovery via programming with large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2404.18400.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(eaay2631), 2020.

Jenny Zhang, Shengran Sheng, Cong Lu, and Jeff Clune. The Darwin Gödel machine: Open-ended evolution of self-improving agents. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.22954.