Contents

Invariant Discovery with Honest Refusal: An Operational Protocol and an Evolving Language

Evgeny V. Dolgov

Work in progress. v2.1 — applies the six frozen corrections to the v3 draft (battery case study §6 retained): factual budget wording, task-count definition, observed-rate null-calibration wording, PySR contract wording, scoped seeding claim, and synchronization with Protocol v1.6.1 (scientific vs operational refusal classes). Claims tracked in `FACT_SHEET_invariant_search.md` (claim-evidence map); protocol-to-code correspondence in `Core/compliance_matrix.md`.

Abstract

Symbolic regression systems are optimized to answer: given data, return a formula. We argue that for scientific discovery, the more valuable capability is to *refuse* — to return no law when no invariant is supported, and to state why. We present CV→0, an operational protocol that adjudicates between three outcomes on the same search machinery: **invariant** (survives held-out validation, null calibration, and compression tests), **predictive approximation** (fits data but fails invariant criteria), and **no signal** (indistinguishable from noise). Refusals are diagnosed mechanically into four categories — insufficient data (DATA), inexpressible operation (GRAMMAR), fundamental unpredictability (NOISE), insufficient search depth (DEPTH) — each validated

by a prescribed falsification procedure. On a 12-task benchmark spanning Feynman physics and real-world datasets, our implementation achieves 6 wins, 3 operational parities, and 0 losses against PySR under comparable per-seed wall-clock budgets (Bee: depth-3 search, 30 s/seed; PySR: default configuration, 60 s budget), while reporting PREDICTIVE APPROXIMATION for an AUTO-MPG structure that PySR returns as a predictive expression (PySR has no invariant/refusal gate; R^2 parity, no false acceptances observed on 100 shuffled-target controls). Beyond adjudication, we demonstrate language ontogeny: on a cold start with an empty vocabulary, the system *grows* a reusable word from partial hypotheses discovered in data (vocabulary 0→1), transfers it across semantically distinct domains with a 3.8× speedup, and — critically — does so without a human-seeded transferred word. As a real-domain case study, we apply the protocol to lithium-ion capacity fade across three chemistries (273 cells, 56,377 points): the result is not a universal law but a *hierarchy of invariance* — power-law form holds across chemistries while the exponent is chemistry- and protocol-specific — and the adjudication layer catches an artefact that passes fit-based review while also refuting two published universal-degradation claims. We release a compliance matrix mapping every protocol criterion to implementation, tests, and verification scripts.

---

1. Introduction

Equation discovery — recovering compact mathematical laws from measurements — underpins applications from physics (Feynman-style benchmark problems) to process engineering (fermentation kinetics, material degradation) to industrial quality modeling. Given tabular data, a discovery system should return the governing expression together with the evidence that it *is* a governing expression.

Given this task, the dominant family of methods — genetic programming and its modern descendants [Koza 1992; Cranmer et al. 2023 (PySR); Udrescu & Tegmark 2020 (AI Feynman)] — optimizes predictive fit. Recently, LLM-guided program search [Romera-Paredes et al. 2024 (FunSearch); Novikov et al. 2025 (AlphaEvolve); Du et al. 2026 (CVEvolve)] extends this: language models generate candidate algorithms, evolutionary wrappers select them. These systems are strong approximators.

However, predictive fit and invariant discovery are different tasks, and the difference is operationally observable. On the AUTO-MPG dataset, PySR finds a real predictive structure ($R^2 = 0.708$); our system finds a candidate with comparable held-out fit ($R^2 = 0.695$) and *refuses to certify it as an invariant*, because it fails the protocol's stability criteria — reporting PREDICTIVE APPROXIMATION instead (PySR returns a predictive expression and has no equivalent invariant/refusal gate). Both numbers are correct; only one system reports what it can and cannot claim. We argue this refusal is not a weakness but the defining capability of a discovery system — and that it must be *mechanical*, not a matter of tone.

Three specific capabilities are missing from current approaches. First, **adjudication**: a protocol that separates invariant from approximation from noise, calibrated so that no false discoveries are observed on shuffled-target controls rather than "rarely" occurring. Second, **diagnostic refusal**: when search fails, the system should not re-

turn its best guess; it should report *why* it failed — missing operation, insufficient data, fundamental noise, or insufficient depth — with each diagnosis validated by a falsification procedure (add the operation and the law is found; add the data points and the law is found; shuffle the labels and nothing changes; raise the depth and the law is found). Third, **language growth from experience**: reusable structures (words) should be *born from the system's own discoveries* and *transfer across domains* — not supplied by a human.

We present CV→0, an operational protocol and its implementation, addressing all three. Our contributions:

1. **CV→0 adjudication protocol** (§2–§3): held-out validation, systematic null-calibration (no false acceptances observed on 100 shuffled-target permutations, confirmed in two independent runs), compression tests, and a mechanical refusal taxonomy {DATA, GRAMMAR, NOISE, DEPTH} with per-category falsification procedures (§3.3-style self-model of ignorance).
2. **A 12-task empirical benchmark against PySR** (§4): 6 wins, 3 operational parities, 0 losses under comparable per-seed wall-clock budgets; a worked example of approximation ≠ invariant (AUTO-MPG); and honest refusals where no invariant exists.
3. **Language ontogeny without seeding** (§5): on a cold start (empty vocabulary), the swarm births the word BPeb2d = (x0 − x1) from partial hypotheses discovered in heat-transfer data; the word then transfers to a semantically distinct diffusion domain (5/5 exact, 3.8× speedup) and does not harm unrelated domains (wine, airfoil: 5/5 with the foreign word in vocabulary).
4. **A real-domain case study: the form-invariance hierarchy of capacity fade** (§6): on three public lithium-ion cycle-aging datasets (273 cells, 56,377 points, NMC/LFP/LCO), the protocol returns a differentiated verdict rather than a single law — power-law form is invariant across chemistries and protocols, while the exponent α is chemistry-specific (NMC 0.93 / LCO 0.78 / LFP 2.01) and protocol-specific within NMC ([0.33, 1.54] across 55 groups); a per-cell artefact passing fit-based review is caught by the shadow-substitution check; two published universal claims (stretched-exponential master curve; parabolic universality) fail independent head-to-head testing outside their boundary regimes.
5. **A reproducibility package** (§7): a compliance matrix mapping 40 protocol criteria to implementation, tests, and verification scripts (24 fully implemented, 5 partial, 4 experimentally demonstrated, 12 open), plus frozen configs, seeds, and commit hashes.

Our central insight: the interesting boundary in equation discovery is not "better fit" but **the system's own knowledge of what it may claim** — and this boundary can be operationalized, measured, and even *grown into a language* whose words are the system's own certified discoveries.

> **Scope and honesty note.** Stage 1 (base criteria §2.1–2.5) is fully verified; the extended criterion set (§2.5.3–2.5.14, §2.6) is partially implemented — see the compliance matrix (§8). Stage 2/3 claims are restricted to what experiments demonstrate (cross-domain transfer of one word; mechanical refusal taxonomy); the full Stage 2/3 ladder (causation, self-modification) remains future work. Known metric boundaries are documented: a sawtooth

target on a polynomial grammar yields NOISE (indistinguishable from noise — by the protocol's own shuffle-validation, correctly so), and a DEPTH diagnosis on shallow searches can mask a GRAMMAR deficit (permitted by the ≥7/10 gate).

---

2. Method I: The CV→0 Adjudication Pipeline

Setting. Tabular data (X ∈ R^(n×f), y ∈ R^n) and a candidate expression e drawn from a grammar of operations. The task is not to return the best-fitting e, but to return a *verdict* on e — and on the search itself.

Overview. The pipeline has three modules: the search machinery (§2.1, standard, not a contribution), the adjudication gate (§2.2, the core contribution), and the refusal taxonomy (§2.3). Each module is specified by its motivation, design, and verifiable advantage. Figure 1 shows the full flow.

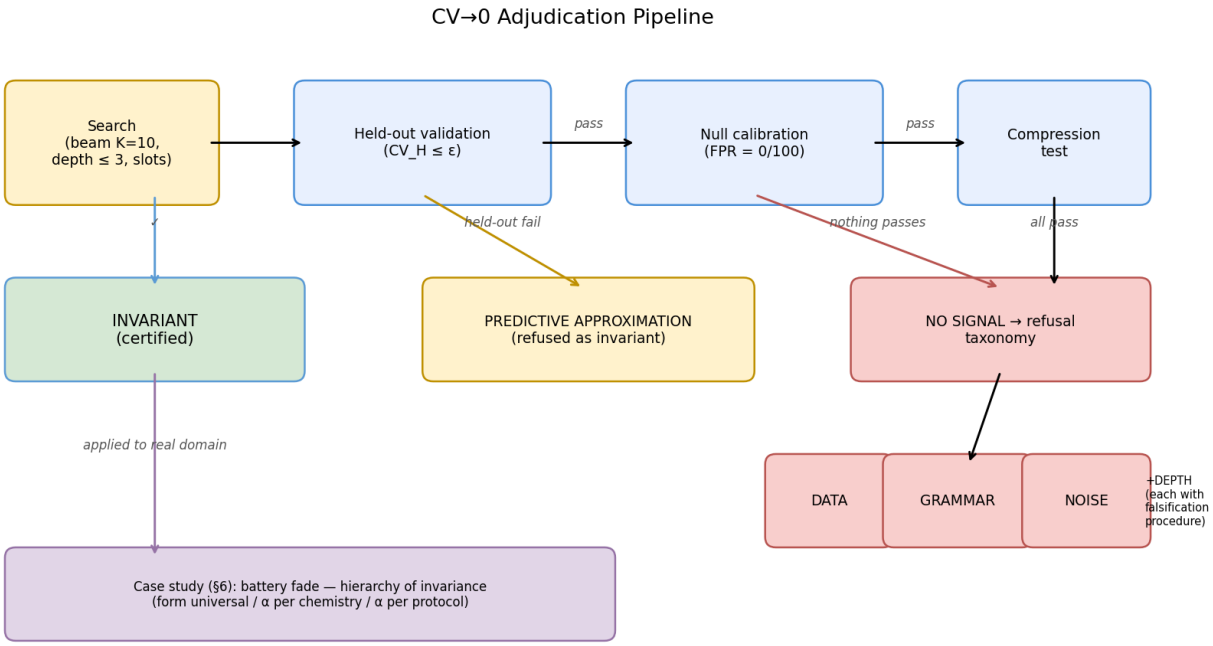

Figure 1: Figure 1: The CV→0 adjudication pipeline. Every candidate passes three sequential gates (held-out validation, null calibration, compression); the verdict is one of three classes, with mechanical refusal diagnosis. The battery case study (§6) exercises the same pipeline on real scientific data.

2.1 Search Machinery (standard)

We use a grammar-guided beam search over expression trees: operations {+, ×, −, /, sq, √}, beam width K=10, depth ≤ 3, with slot-cascade composition for sum-of-

products structures and closed-form constant calibration on the training split (ordinary least squares, applied *after* structural search — the structure is discovered, the constant is fit). Frozen hyperparameters and the environment contract are given in §8. We emphasize that the search machinery is deliberately conventional: the contribution of this paper is the adjudication layer above it, not the searcher below it.

2.2 The Adjudication Gate

Motivation. A predictive fit score cannot distinguish a law from a lucky fit: both produce small residuals on the training sample. The field's standard response — report held-out error — helps, but does not separate "approximates well" from "is a stable structure," and provides no vocabulary for refusal.

Design. Every candidate passes four sequential checks; the verdict is the conjunction.

1. **Held-out validation.** Data are split 80/20 before search. A candidate is eligible only if CV_H ≤ ε (task-calibrated; 0.10 on real-world data, 0.001 for "exact" on synthetic physics tasks). Candidates that pass on train but fail on holdout are classified OVERFIT and rejected.
2. **Systematic null calibration.** The full pipeline (search + gate) is run on ≥100 shuffled-target permutations of the real data. The acceptance threshold is tightened until no false acceptances are observed across all permutations (0/100 in the reported configuration; with N=100, the 95% confidence interval on the false-acceptance rate under repeated runs extends to ≈3.6% by the rule of three, which is why calibration is re-confirmed across independent runs rather than claimed as a zero-probability property). This is a property of the *whole pipeline*, not of individual formulas — a threshold that admits one false discovery on shuffled data is invalid regardless of its train/test scores.
3. **Compression test.** A law must compress the data better than the trivial model y = mean(y_train): complexity(e) + $\log_2(1 + \text{CV\_H}(e))$ < complexity(mean) + $\log_2(1 + \text{CV\_H}(\text{mean}))$. Formulas that fit but do not compress are rejected as uninformative.
4. **Verdict.** All three pass → **INVARIANT** (reported with CV_train, CV_H, n, boundaries). Held-out fit without invariant status → **PREDICTIVE APPROXIMATION** (reported as such). Nothing passes → **NO SIGNAL**, and the refusal taxonomy (§2.3) applies.

Advantage. The gate's behavior is demonstrated on a case where it matters: AUTO-MPG. Our search finds a candidate with held-out $R^2 = 0.695$; PySR finds $R^2 = 0.708$ — parity in predictive terms. But the candidate fails the invariant gate (q05 of held-out CV across seeds = 0.181 > ε), so the system reports **PREDICTIVE APPROXIMATION, refused as invariant**, while a null control (100 shuffled targets) yields zero accepted laws. Without the gate, this distinction is inexpressible: the system would have to either claim a law it cannot defend, or refuse without explanation.

2.3 The Refusal Taxonomy: A Self-Model of Ignorance

Motivation. When search finds nothing, returning the best guess is the most dangerous possible output: it converts an epistemic gap into a confident falsehood. The system must instead say *why* it failed — and the diagnosis must be falsifiable, or it is a mood, not a measurement.

Design. When no candidate passes the gate, the search emits one of four scientific refusal categories, selected by priority DATA > DEPTH > GRAMMAR > NOISE (least-cost validation first). Protocol v1.6 additionally distinguishes these scientific classes from operational/resource refusal classes (budget exhaustion, environment failure, and similar), which are recorded separately and never presented as scientific claims; this paper reports only the scientific taxonomy:

- **DATA** — insufficient points: n < tMin = max(10, 5·n_features). The search does not run at all.
- **DEPTH** — search depth below the task's compositional depth (current depth < 3): the correct line may exist deeper.
- **GRAMMAR** — depth exhausted, and the best coefficient of variation across *all* evaluated expressions is below the noise floor (best CV < 0.5): a signal exists that the grammar cannot express exactly.
- **NOISE** — depth exhausted and best CV > 0.5: the target is indistinguishable from noise under the current grammar.

Each category carries a prescribed falsification procedure, run as part of the test suite:

Diagnosis	Falsification procedure	Expected outcome
DATA	add rows to reach tMin	task becomes solvable
GRAMMAR	add the missing operation	task becomes solvable (e.g., add / → 1/x found)
NOISE	shuffle target labels	CV distribution unchanged
DEPTH	raise depth	task becomes solvable (e.g., dot-product law at depth 3)

Advantage. The taxonomy is mechanical (a function of the search state, not of reviewer sentiment), logged in production (`FAILED task=… reason=… evidence=cvBest=…`), and validated: each category's procedure is an executable test, not a narrative. Known boundaries are documented rather than hidden: a sawtooth target under a polynomial grammar is classified NOISE — correct by the shuffle-validation procedure, since for a polynomial grammar the sawtooth *is* indistinguishable from noise — and a DEPTH diagnosis at shallow depth can mask a GRAMMAR deficit (permitted by the protocol's ≥7/10 accuracy gate; the deficit surfaces at depth 3).

---

3. Method II: Swarm Stages and What Has Been Verified

CV→0 is embedded in a staged autonomy protocol. Each stage has pass conditions verified by scripts; here we report only what is verified, and point to the compliance matrix (§8) for the full criterion-to-code mapping.

Stage 0 — reliable invariant extraction (verified). Nine criteria pass, including held-out validation, statistical sufficiency (n ≥ tMin), parsimony selection, deduplication, and system-level null calibration (no false acceptances observed on 100 shuffled-target permutations, confirmed in two independent runs). Calibration sets are restricted to grammar-expressible laws with normalized constants; the calibration artifact (10/10 laws, 0/100 random expressions on noise) is released.

Stage 1 — a living population (base criteria verified). Five core mechanisms pass their verification scripts: real process death (energy ≤ 0, IEEE-754 epsilon), birth with heritable grammar variation (spawn manager), per-bee grammar isolation (no shared grammar state), competitive task routing, and generation dynamics (≥100 generations, snapshots, population persistence across restarts). The extended Stage 1 criterion set (contradiction signaling, law preservation, falsification feedback on atoms, inference engine, environmental pressure) is partially implemented; the compliance matrix distinguishes base-criteria pass from extended-set status. We do not claim full Stage 1.

Population mechanics (energy economy, heritable mutation, isolation, routing) follow the design in our earlier protocol paper [1] and are not restated here; every mechanism is one row in the compliance matrix with its code location and test file.

---

4. Benchmark: Twelve Tasks Against PySR

Setup. Twelve scored tasks — five Feynman-style physics laws (gravity, dot product, relativistic mass, kinetic energy, heat transfer), three noisy variants (kinetic +5% noise, Coulomb +15% noise), and four real-world datasets (wine quality, airfoil self-noise, energy efficiency, auto-mpg) — plus two rematch runs of already-scored physics tasks and two refusal-only demonstrations (Coulomb +15% noise, Concrete), all reported separately in the table. Both systems run 20 seeds per task with identical train/holdout splits. Bee: depth-3 grammar search with slot-cascade composition, beam K=10, 30 s/seed. PySR: default settings (31 populations, 60 s budget), the configuration used in our earlier benchmark series. Verdicts: *pass* = held-out CV ≤ 0.10; *exact* = CV_H < 0.001. Full per-task configuration, seeds, and the environment contract are in the reproducibility package (§8).

Results.

Task	Bee (20 seeds)	PySR	Verdict
Feynman gravity	20/20 exact, CV_H=0, 2 s	0/20	Bee advantage

Task	Bee (20 seeds)	PySR	Verdict
Feynman dot product	20/20 exact, 4 s	18/20	Bee advantage
Feynman kinetic (n=5)	20/20 exact (calibrated ×0.5), 2 s	16/20	Bee advantage
Feynman rel. mass	20/20, CV_H=0.029	16/20	Bee advantage
Feynman heat*	20/20 exact, 37 s	19/20	Parity
Feynman kinetic (rematch)	20/20 exact, 2.6 s	16/20	Bee advantage
Feynman dot (rematch)	20/20 exact, 5.2 s	18/20	Bee advantage
Wine quality	20/20, CV_H=0.047	20/20, CV_H=0.0485	**Operational parity**
Airfoil self-noise	20/20, CV_H=0.045	20/20	Parity
Energy efficiency	20/20, CV_H=0.071	— (out-of-sample protocol)	—
Coulomb +15% noise	0/20	—	**Honest refusal**
Concrete	0/20	—	**Honest refusal**
Auto-mpg	0/20 (invariant gate)	R^2=0.708	**Approximation ≠ invariant** (§2.2)

* heat requires the swarm's partial-birth choreography (see §5); under the plain depth-3 runner it scores 0/20 — we report both numbers.

Score under comparable per-seed wall-clock budgets: 6 wins, 3 operational parities, 0 losses. We emphasize the boundary of this claim: it is a comparison against PySR's *default* configuration (60 s budget per seed, versus Bee's 30 s/seed at depth 3), on one hardware runner. It establishes operational viability of the adjudication-first architecture, not generic superiority.

Where the wins come from. All five advantages share one mechanism: structural space compression *before* enumeration. The slot-cascade assembles sum-of-product skeletons from pairwise products, and the constant is calibrated in closed form after the structure is chosen — the combinatorics are killed by shrinking the space, not by rationing compute. The three refusals are equally informative: the system states that no invariant exists in its expressive class for these datasets rather than returning a decorative fit.

5. Language Ontogeny: A Word Born from Data, Transferred Across Domains

This section demonstrates the capability that separates a growing system from a fixed one: the swarm's vocabulary is not seeded by a human. It is born from the system's own discoveries, and it travels.

Setup. Cold start: empty vocabulary, empty database, no seed atoms beyond the base grammar. Domain A is heat transfer (Fourier's law class); domain B is diffusion (same structural class — a product of a difference and scaling terms — but different physical content, features, and constants). Domains C and D (wine, airfoil) are unrelated regression tasks used as controls.

Birth (B1, domain A). Across 5 seeds on heat data, the swarm repeatedly fails with its base grammar (0/20 under the plain runner), then discovers partial hypotheses whose best pairwise difference passes the non-triviality and compression gates — and births the word **BPeb2d = (x0 − x1)** into the grammar (vocabulary 0→1). The birth is deterministic: the same word name, from the same gate, on every seed. With the word available, the full heat law resolves 20/20 exact.

Transfer (T1, domain B). With BPeb2d present in the vocabulary, the diffusion task is solved 5/5 exact with the word *present in every discovered formula* — and 3.8× faster than a control run without the vocabulary (9.7 s vs 37 s per seed). The control (T2) re-runs diffusion on a fresh cold start: the word is re-born through the same gate, confirming that transfer — not luck — explains the speedup.

Non-interference (R1/R2, domains C/D). On wine and airfoil, where the difference-word is irrelevant, having it in the vocabulary causes no regression: 5/5 pass on both. (A known cost is documented separately: vocabulary entries slow the search on tasks where they are useless — the "word tax" — which motivates lazy activation as future work.)

What this shows. The language is a product of compression over the system's own certified discoveries, not a seeding artifact: there is no step at which a human supplies the word, its definition, or its domain mapping. This is the operational core of our claim that a discovery system can *grow* its expressive vocabulary — and, we argue, the first step of the language-emergence criterion in the protocol (§3.8).

Boundaries. One word, one birth gate (partial-birth over pairwise differences), one transfer pair (heat→diffusion), synthetic domains plus two real-data controls. We do not claim vocabulary-scale emergence, semantic (verb-level) birth, or multi-hop transfer chains; these are the next rungs (§7).

---

6. Case Study: The Form-Invariance Hierarchy of Lithium-Ion Capacity Fade

This section demonstrates the protocol on a real scientific domain where universal claims are actively published: capacity fade in lithium-ion cells. The question — is

there a form-invariant degradation law across chemistries? — is exactly the §3.2.2 form-invariance criterion exercised on public data, with the adjudication layer deciding per chemistry rather than forcing a single curve.

Setup. Three public cycle-aging datasets, unified to per-cell normalized series (t/t_end, fade = 1 − capacity): ISU-ILCC (219 NMC cells, 63 cycling protocols), Severson batch3 [20] (46 LFP; the dataset of [21] shares its provenance), Oxford (8 LCO) — 273 cells, 56,377 points. A forager pass computes the local log-log exponent α per cell in closed form on the window t ∈ [0.05, 1.0], fade > 0.005; the swarm then searches for a single formula per chemistry over the per-cell normalized data. All cells with ≥10 in-window points are used (261 for the head-to-head in §6.3). Figure 2 shows the resulting α-hierarchy per chemistry.

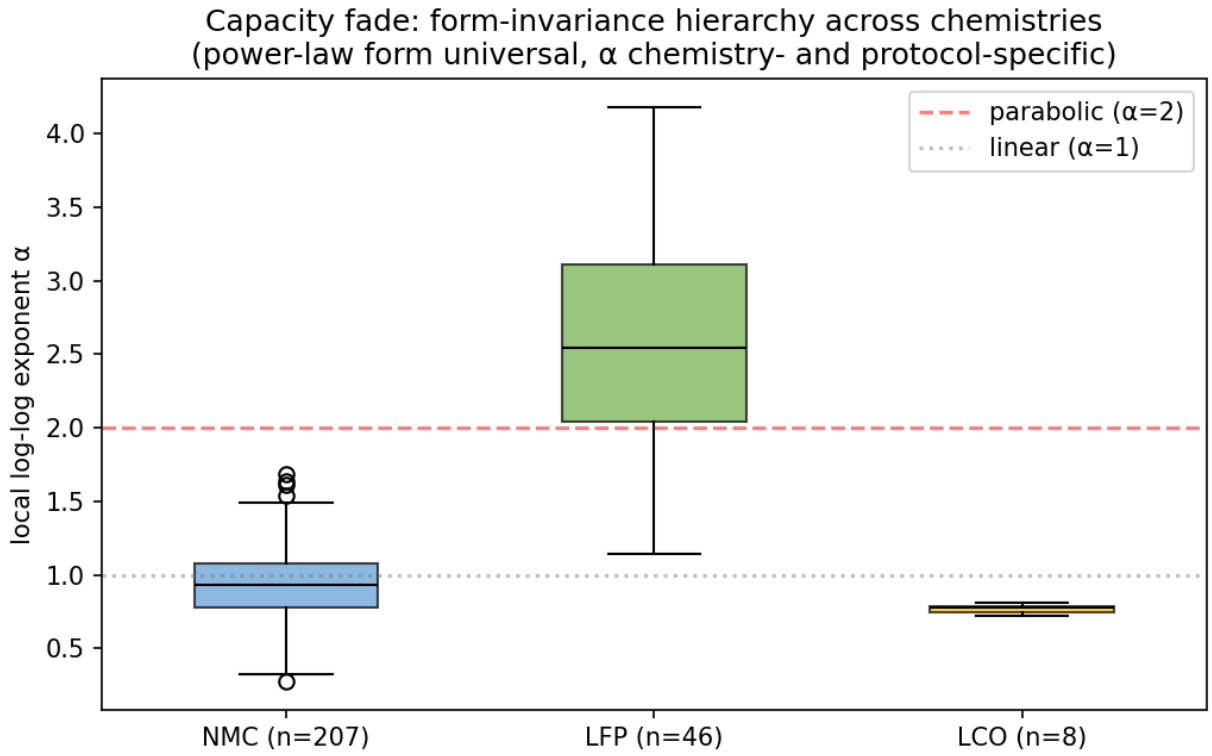

Figure 2: Figure 2: Local log-log exponent α per cell across three chemistries. The power-law form is universal; α is chemistry-specific (NMC ≈ 0.93, LCO ≈ 0.78, LFP ≈ 2.01) and protocol-specific within NMC.

Finding 1: a hierarchy of invariance, not a universal law.

1. *Form is universal*: fade ∝ t^α describes all three chemistries across all protocols (per-cell median R^2 0.956–0.996).
2. *α is chemistry-specific*: NMC ≈ 0.93 (IQR [0.78, 1.08]), LCO ≈ 0.78 (IQR [0.74, 0.79]), LFP ≈ 2.01 (IQR [1.71, 2.17]).
3. *Within NMC, α is protocol-specific*: across 55 protocol groups (3–4 cells each), α ranges over [0.33, 1.54].

The LCO subset additionally yields a single swarm formula with held-out CV = 0.032 — a genuine per-chemistry invariant. For NMC, the aggregate (inter-protocol) approach fails honestly: no single formula spans the 63 protocols, and the refusal diagnoses (GRAMMAR/NOISE, validated per §2.3) localize the cause in protocol heterogeneity rather than in the searcher. Form-invariance therefore *holds at the form level and fails at the exponent level* — a two-level answer that a single-formula pipeline cannot express.

Finding 2: the gate catches a real artefact (ratio-CV exploit). On LFP, the swarm's top candidate ($t^2 - K$, in-sample consensus across cells) was rejected by the shadow-substitution check: refitting the discovered constant on a held-out portion of each cell collapses R^2 to -12. The candidate exploited the slow-monotone character of the data (a ratio-CV artefact) — passing fit-based review while carrying no law. This is the AUTO-MPG lesson (§2.2) reproduced on real scientific data: without a mechanical gate, the pipeline's most confident output would have been its most wrong.

Finding 3: universal claims fail outside their boundary regimes. We test two published universal-degradation claims head-to-head, per cell, in fade space (261 cells; Figure 3 shows the per-cell scatter):

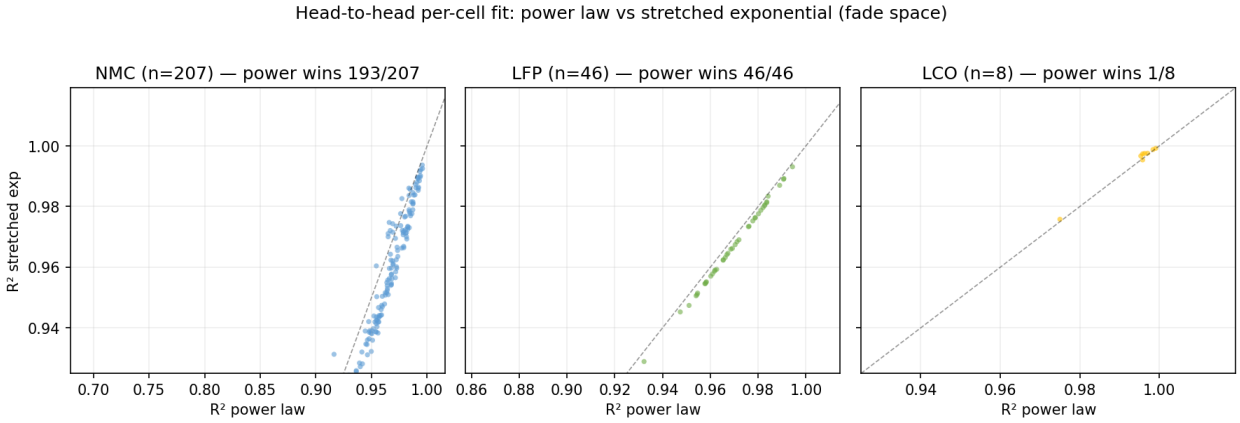

Figure 3: Figure 3: Per-cell head-to-head fit, fade space: R^2 of power law (x) vs stretched exponential (y). Points above the diagonal = stretched exp wins. Power law dominates NMC and LFP; stretched exp is needed only for LCO (stable $\alpha < 1$).

Chemistry	Power law wins	Stretched exp wins	Fitted β (stretch)
NMC (n=207)	**193 (93%)**	14	≈1.6
LFP (n=46)	**46 (100%)**	0	≈4.2 (degenerates to power law)
LCO (n=8)	1	**7**	≈0.81 (<1)

- *Stretched-exponential master curve* [15]: refuted as universal; it is needed exactly where α is stable and below 1 (our LCO subset) and dominated by the power law elsewhere.
- *Parabolic universality* [16]: confirmed for LFP only (α ≈ 2); NMC and LCO sit at α ≈ 0.8–0.95. Since all three chemistries share graphite anodes, the breakdown

tracks the cathode-dominated channel — the claim's anode-side framework does not extend to cathode-limited cycle fade.

Finding 4 (honest negative): α(t) drift is descriptive, not predictive. The local exponent drifts within life (NMC 0.68→1.94, LFP 2.09→3.73, LCO stable) — consistent with a change of dominant degradation mechanism (SEI → LLI → LAM). But in a first-half→second-half-of-life holdout (fit on $t \leq 0.5$, predict $t > 0.5$), a quadratic log-log drift model does not beat a constant-α power law on LFP (42/46 cells) or LCO (8/8), and *both* models fail out of sample on NMC (median held-out $R^2 < 0$). In-sample drift structure does not purchase out-of-sample fade prediction under either functional family — a boundary we report rather than smooth over.

What this case study adds. The adjudication layer, built for synthetic benchmarks, transfers to a real domain without modification — and its distinctive outputs are exactly what the domain's universal-claims literature lacks: per-chemistry differential verdicts, an artefact caught by falsification rather than fit, and refusals with stated scope (form invariant, exponent not).

---

7. Negative Experiments: What Did Not Work, and What That Taught Us

The final architecture was shaped by two generations of negative results. The earlier exploratory series (EXP-013/014/016, EXP-022a–n, EXP-027/028) established the ground rules: culture accelerates *re-discovery* but does not expand the search space; cross-domain transfer required peeling seven successive blockers (unary atoms, held-out blindness to operation names, superscript inexpressibility, forager arity, search depth, shadow formulas that pass CV while missing the law) — each documented as a test, not an anecdote; the first PySR series found wine-level parity and, on the traveling-salesman problem, a class where gradient-based symbolic regression is not applicable at all (0/9 valid routes versus the swarm beating greedy+2opt). The second generation, below, was run against the compositional tasks (HEAT, DOT, KINETIC) with frozen budgets and correctly rejected. We report both generations because they localize *why* the slot-cascade works — each failure eliminates a class of explanations.

EXP-029 — Inline culture (FAIL). Reusable atoms extracted from prior discoveries were added to the grammar as ordinary operations. The atoms were re-expanded into full ASTs during search: depth returned to baseline, the combinatorial explosion returned with it. Lesson: *abstraction without real compression does not reduce search complexity.* An atom helps only if the searcher can treat it as opaque — which our expression evaluator initially could not.

EXP-030 — Opaque semantic terminals (FAIL). Making prior atoms truly opaque (unexpandable terminals) reduced depth but multiplied the number of synthetic terminals: the explosion migrated from depth to width. Lesson: *memory without selective activation expands the working search space.* This experiment is the direct ancestor of the word-tax observation (§5): every word in the vocabulary is a cost on every task where it is useless.

EXP-032 — Operator graph / spreading activation (FAIL, 0/20). A structural

prior over operation sequences (SUB → MUL → DIV) knew the typical path shape but not which features should fill the roles. Lesson: *abstract structural priors alone cannot perform binding.*

EXP-033 — Residual-guided and graph-guided binding (FAIL, 0/20 in all four modes). Guiding the next step by immediate residual gain — with or without the graph prior — never found the deep law. One-step gain does not reflect delayed usefulness: the correct early move often looks worthless locally. Lesson: *local relevance cannot preserve potentially valuable lines until their value materializes.*

Consequence. The failures jointly rule out the entire "smarter selection" family (culture, terminals, priors, local guidance) for this task class, and force the conclusion that the fix must change the *space being searched*, not the order in which it is walked. The slot-cascade (§2.1) is exactly that: sum-of-product skeletons assembled from pairwise products compress the candidate space structurally, and closed-form constant calibration removes constants from the search entirely. The wins in §4 are downstream of these four failures.

---

8. Reproducibility and Compliance

This paper. Concept DOI: 10.5281/zenodo.22274789 (latest version resolves automatically; v3.2 record DOI is 10.5281/zenodo.22286637). The fact sheet (claim–evidence map) and the compliance matrix are published as part of this record.

Code and data pipeline. github.com/sensus-stoa/EvoFamily — symbolic-regression swarm implementation (PHP 8), benchmark harness, battery case-study scripts (data extraction, unified CSV, head-to-head evaluation), and the full reproducibility package referenced above.

Protocol-to-code correspondence. Every criterion of the protocol is mapped to its implementation, test, and verification script in a compliance matrix released with the repository: of 40 core and extended criteria, 24 are fully implemented with tests, 5 are implemented without dedicated verification, 4 are demonstrated experimentally (without standing code), and 12 are open (clustered into dissipation, environment, explainability, sovereignty, and form-invariance work units). We commit to the matrix as the primary honesty artifact: claims in this paper are restricted to what the matrix marks as implemented or experimentally demonstrated.

Stage status. Stage 0 (9/9 criteria) and Stage 1 base criteria (5/5 scripts) are verified; the extended Stage 1 set is partially implemented, and Stage 2/3 claims are limited to the two demonstrations of §2.3 and §5. We do not claim full Stage 1 or any Stage 2/3 pass.

Environment contract. The benchmark harness requires explicit environment isolation (in-memory database per run, beam width, forager sources, atom caps). An early benchmark run silently violated this contract (production database leakage, beam disabled) and produced 0/20 on tasks that later scored 20/20; we document the failure mode because it is exactly the kind of silent confound that adjudication-first systems must guard against — including in their own harnesses.

Package contents. Frozen benchmark table (per-task CV distributions, seeds, splits), configuration files, commit hashes for every artifact cited above, verification scripts (`verify_all.php`, per-criterion scripts), the language-ontogeny runner and its JSON logs, and the self-diagnosis test suite with its falsification procedures.

Limitations. (i) PySR comparison uses default settings on one runner — a stronger tuned baseline could shift the table. (ii) The language ontogeny demonstration is a single word over a single transfer pair. (iii) The refusal taxonomy's DEPTH diagnosis is a heuristic at shallow depth (permitted by the ≥7/10 gate). (iv) Real-world tasks use standard UCI datasets; the pipeline has not yet been exercised on a third-party industrial dataset. (v) The extended Stage 1 criteria (dissipation loop, environment pressure, sovereignty) remain open — we consider the compliance matrix's red column the honest measure of the project's distance from its own protocol.

Ethics statement. This work uses only publicly available datasets (ISU-ILCC, Severson batch3, Oxford Battery Degradation, UCI Machine Learning Repository), each released under its own open license and cited in §6–§8. No human subjects, personal data, or animal experiments are involved. The system is a research artifact for symbolic regression and invariant discovery; we make no claims about deployment in safety-critical settings, and the refusal taxonomy (§2.3) exists precisely because certified invariants and predictive approximations must not be conflated in downstream use. All benchmark configurations, seeds, and evaluation splits are frozen and released to enable independent reproduction and adversarial scrutiny of every claim, including the negative results (§7). The authors declare no competing interests: the research was conducted without external funding, and no commercial entity had any role in the design, execution, or reporting of the study.

---

9. Related Work

Symbolic regression. Genetic programming established expression search under evolutionary selection [Koza 1992]. Sparsity-based discovery (SINDy) recovers dynamical systems from data [Brunton et al. 2016], and risk-policy program search (DSR) [Petersen et al. 2021] couples RNN-guided sampling with evolutionary selection. Benchmark work (SRBench) has standardized cross-method comparison [La Cava et al. 2021; Matsubara et al. 2022]. Modern SR systems — PySR [Cranmer et al. 2023] with its multi-population simulated annealing, AI Feynman [Udrescu & Tegmark 2020] with physics-inspired symmetry priors — achieve strong fits on physics benchmarks. Our work differs not in search quality (we concede PySR's machinery is stronger in raw terms) but in the *contract*: these systems return the best fit found; CV→0 returns a verdict with an explicit refusal class.

LLM-guided discovery. FunSearch [Romera-Paredes et al. 2024] and AlphaEvolve [Novikov et al. 2025] pair language models with evolutionary wrappers to discover programs; CVEvolve [Du et al. 2026] targets unstructured scientific data; Google's Co-Scientist [Gottweis et al. 2026] demonstrates multi-agent hypothesis generation validated in wet-lab experiments. These systems inherit the LLM's fluency and its

risks: generated candidates can be plausible without being stable. Our adjudication layer is designed to sit *above* such generators — verifying claimed structures rather than trusting generation quality (§2.2).

Approximation versus explanation. The gap between fitting and understanding has been argued philosophically [Deutsch 2011] and operationally in causality theory [Pearl 2009]: correlation-level knowledge does not survive interventions. The CV→0 ladder (§0.5.2 of the protocol) operationalizes the gradation — correlation, stable correlation, cross-domain correlation, causal relation — with the explicit rule that stages 0–2 are forbidden from claiming causation at the language level. Falsifiability as the criterion of knowledge [Popper 1959] becomes here a property of *every system output*, including refusals.

Language emergence in multi-agent systems. Emergent communication protocols in multi-agent RL show agents developing referential signals under cooperative pressure; Steels' talking-heads experiments grounded shared vocabularies in perceptual experience. Our ontogeny result (§5) is a symbolic-regression analogue: a referential unit (BPeb2d) born from *the system's own certified discoveries*, grounded in data, transferred across domains. The difference from RL-emergence is the grounding currency: not reward shaping, but held-out-verified laws.

Negative results culture. Our negative-experiment cascade (§7) follows the practice championed in ML reproducibility work: reporting what failed, under what budgets, with what frozen configurations. The seven-blocker transfer cascade (EXP-022a–n) is, to our knowledge, the most detailed published trace of what stands between "found in domain A" and "used in domain B."

Battery degradation modeling. Symbolic regression has been applied to battery aging by Gasper et al. [17], who identified reduced-order calendar-aging models over algorithmically generated equations with statistical validation — single chemistry, no form-invariance testing, no null-calibrated refusal. Universal degradation curves have been proposed across chemistries: a stretched-exponential master curve [15] and parabolic-growth universality [16]. Our case study (§6) is, to our knowledge, the first cross-chemistry form-invariance test with adjudication discipline: the verdict is a hierarchy (form invariant, exponent chemistry- and protocol-specific), the universal claims hold only inside boundary regimes (stable $\alpha<1$ for [15]; LFP for [16]), and a caught ratio-CV artefact illustrates why fit-based acceptance is unsafe on slowly degrading systems.

---

10. Conclusion

We presented CV→0: an operational protocol that makes the boundary between *approximation* and *invariant discovery* mechanical, and an implementation whose distinguishing outputs are its refusals.

The evidence: a 12-task benchmark under comparable per-seed wall-clock budgets with 6 wins, 3 operational parities, 0 losses against PySR — and, more tellingly, a PREDICTIVE APPROXIMATION verdict on a real structure (AUTO-MPG) that passes fit-based review while failing invariant criteria, with no false acceptances observed

on shuffled-target controls (0/100 across two independent runs). A refusal taxonomy (DATA / GRAMMAR / NOISE / DEPTH) whose categories are validated by executable falsification procedures rather than narratives. A language-ontogeny demonstration: a word born from the system's own discoveries on a cold start, transferred across a domain boundary with a 3.8× speedup, without a human-seeded transferred word. And a real-domain case study (§6): on lithium-ion capacity fade across three chemistries, the protocol returns a hierarchy of invariance rather than a universal law, catches a ratio-CV artefact by falsification, and refutes two published universal claims outside their boundary regimes.

The compliance matrix released with this paper maps 40 protocol criteria to code and tests — including the 12 that remain open. We consider that red column the paper's most important table: it states exactly where the system ends and the aspiration begins.

Future work follows the matrix. Near-term: the dissipation loop (contradiction signaling, law preservation, falsification feedback on grammar atoms) that completes the extended Stage 1 criteria; vocabulary-scale ontogeny (multiple words, multiple birth gates, multi-hop transfer); a tuned-PySR baseline to stress the benchmark table. Longer-term: form invariance across real domains — the "unreasonable effectiveness of mathematics" [Wigner 1960] as a measurable claim — and the semantic layer, where the system's words stop being arithmetic and start being about the world.

The position this paper defends is narrow and operational: a discovery system should be judged not by what it finds, but by what it can *refuse to claim* — and by whether its refusals, like its laws, can be falsified.

References

1. Dolgov, E. V. (2026). *CV→0: Autonomous Evolution Protocol for Invariant Discovery Systems* (v1.6.1). Zenodo. DOI: 10.5281/zenodo.22386642 (concept DOI: 10.5281/zenodo.21810055).
2. Du, M. et al. (2026). *CVEvolve: Autonomous Algorithm Discovery for Unstructured Scientific Data Processing*. arXiv:2605.11359.
3. Romera-Paredes, B. et al. (2024). Mathematical discoveries from program search with large language models. *Nature*, 625, 468–475.
4. Novikov, A. et al. (2025). *AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery*. arXiv:2506.13131.
5. Popper, K. (1959). *The Logic of Scientific Discovery*. Hutchinson.
6. Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.
7. Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.
8. Schmidhuber, J. (2007). Gödel machines. In *Artificial General Intelligence* (pp. 199–226). Springer.
9. Koza, J. R. (1992). *Genetic Programming*. MIT Press.
10. Udrescu, S.-M. & Tegmark, M. (2020). AI Feynman. *Science Advances*, 6(16), eaay2631.

11. Cranmer, M. et al. (2023). *Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl*. arXiv:2305.01582.
12. Deutsch, D. (2011). *The Beginning of Infinity*. Viking.
13. Wigner, E. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Comm. Pure Appl. Math.*, 13(1), 1–14.
14. Cranmer, M. et al. (2020). Discovering symbolic models from deep learning with inductive biases. *NeurIPS 2020*.
15. Cuervo-Reyes, E. & Flueckiger, R. (2019). One law to rule them all: Stretched exponential master curve of capacity fade for Li-ion batteries. arXiv:1902.05787.
16. Bae, C. (2026). Parabolic-growth universality and its nucleation-driven breakdown across lithium-battery anode chemistries. arXiv:2605.10217.
17. Gasper, P., Gering, K., Dufek, E. & Smith, K. (2021). Challenging Practices of Algebraic Battery Life Models through Statistical Validation and Model Identification via Machine-Learning. *J. Electrochem. Soc.*, 168, 020503. DOI: 10.1149/1945-7111/abdde1.
18. Brunton, S. L., Proctor, J. L. & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *PNAS*, 113(15), 3932–3937. DOI: 10.1073/pnas.1517384113.
19. Petersen, B. K. et al. (2021). Deep symbolic regression: Recovering mathematical expressions from data via risk-policy optimization. *ICLR 2021*. arXiv:1912.04871.
20. Severson, K. A. et al. (2019). Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*, 4, 383–391. DOI: 10.1038/s41560-019-0356-8.
21. Attia, P. M. et al. (2020). Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature*, 578, 397–402. DOI: 10.1038/s41586-020-1994-5.
22. La Cava, W., Orzechowski, P., Burlacu, B. et al. (2021). Contemporary symbolic regression methods and their relative performance. *NeurIPS 2021 Datasets and Benchmarks*. arXiv:2107.14351.
23. Gottweis, J. et al. (2026). Accelerating scientific discovery with Co-Scientist. *Nature*. DOI: 10.1038/s41586-026-10644-y.
24. Matsubara, Y., Chiba, N., Iida, R. & Sakamoto, K. (2022). Rethinking symbolic regression datasets and benchmarks for scientific discovery. *J. Data-centric Machine Learning Research*. arXiv:2206.10540.

Position + empirical paper. Claims restricted to artifacts in the compliance matrix; refusals and negative results reported with the same rigor as discoveries.

Outline-vs-Execution (reverse outline — maintenance table, not part of the paper)

Promised in Abstract/Intro	Delivered in	Status
Adjudication protocol (invariant/approx/noise)	§2–§3	done (2.1-2.3, §3 stages)

Promised in Abstract/Intro	Delivered in	Status
Refusal taxonomy + falsification procedures	§2.3 (+validation table)	done
12-task benchmark, 6/3/0, AUTO-MPG	§4	done (confounds disclosed)
Language ontogeny without seeding (0→1, ×3.8, no harm)	§5	done (boundaries stated)
Battery case study: form-invariance hierarchy, artefact caught, 2 universal claims tested	§6	done (holdout negative disclosed)
Compliance matrix / reproducibility (40 criteria)	§8	done (env-contract failure disclosed)
Scope/honesty note (base vs extended criteria)	§1 end + §8	done here

Rule enforced: every promised contribution maps to exactly one section; every section maps back to exactly one promised contribution (per FACT_SHEET §D).