The Cross-Universe Symbolic Regression Tournament: Survival of the Fittest Laws

Joshua Kyan Aalampour

Kyoto, Japan

May 12, 2025

Abstract

We introduce the Cross-Universe Symbolic Regression Tournament (CU-SRT), a domain-agnostic, Darwinian alternative to first-principles derivation for discovering cross-dataset closed-form laws. Inspired by natural selection, the method treats overfitting not as a vice to be avoided but rather as a mutation-generating force. Symbolic regression engines deliberately overfit candidate formulas to individual "universes" (independent data realms), populating a diverse hypothesis pool. These candidates are then subjected to evolutionary pressure by competing for survival based on cross-universe generalization accuracy and parsimony. Only the fittest equations endure, crystallizing into high-confidence conjectures of the underlying scientific law. CU-SRT thus tempers empirical noise in a crucible for symbolic law discovery (a dancing star born of data-driven tournament, not dogma!) Optional modules include heteroskedastic universe weighting, stochastic grammar annealing, causal-graph pruning, and Bayesian tournament scoring. These evolve as plug-in mutations within the tournament ecosystem, integrating seamlessly while preserving all theoretical guarantees.

$$\boxed{\mathcal{L}^{\star} = \arg\max_{\varphi\in\mathcal{F}} \big\{ \bar{G}(\varphi) - \lambda\,\ell(\varphi) \big\}}$$

The Law of Cross-Universe Survival: Darwinian Optimization in Algebraic Form

Contents

List of Symbols

$\mathcal{U}$, u	Set of universes, single universe index
$X^{(u)}, Y^{(u)}$	Features and targets in universe u
$\mathcal{L}$	Hidden ground-truth law to be discovered
$\mathcal{L}^{\star}$	Tournament champion (estimator of $\mathcal{L}$)
φ	Candidate symbolic formula
$\mathcal{G}_{k,D}$	Grammar (k primitives, depth D)
$\mathcal{F}$	Hypothesis space (all expressions in $\mathcal{G}_{k,D}$)
$C_{k,D}$	Grammar size bound k^D
$\ell(\varphi)$	Symbol count (description length) of φ
$G_u(\varphi)$	Generalization score on universe u (clipped to $[0, 1]$)
$\bar{G}(\varphi)$	Cross-universe mean score
$\mathcal{T}(\varphi)$	Tournament score $\bar{G} - \lambda\ell$
γ, λ, τ	γ in-sample cut; λ complexity weight; τ survival bar
N, T_u	N universes; T_u samples in universe u
$R_u(\varphi)$	Population risk in universe u
η	Additive noise term
ν^2	Upper bound on residual variance
$\mathcal{C}$	Global candidate pool (mutation set)
$\mathcal{S}$	Survivor set after tournament
w_φ	Ensemble weight of survivor φ
W	Total ensemble weight $\sum_{\varphi \in \mathcal{S}} w_\varphi$
p_φ	Normalized ensemble weight w_φ / W
$\mathcal{L}_{\text{ens}}$	Ensemble law (weighted survivor average)

1 Introduction: Derivation by Tournament

At its core, the Cross-Universe Symbolic Regression Tournament (**CU-SRT**) (pronounced "cursed" because it's cooler that way) is a tournament-style derivation mechanism designed for domains where first-principles analysis is inaccessible, unknown, or too complex to execute analytically. The method is based on a key philosophical inversion: instead of treating overfitting as a vice, we embrace it as a generative force. The goal is not to guess the true law directly, but to let it emerge as the sole survivor of a rigorous, adversarial elimination process carried out across many data "universes".

1.1 Outline

1. *Overfitting is deliberate and local.* Within each universe (e.g. an asset, experiment, or physical system), we allow symbolic regression to overfit, capturing both genuine structure and idiosyncratic noise. These locally accurate formulas (which we shall call our "candidates") create a rich hypothesis pool that we will exploit in later phases.

2. *Universality is enforced globally.* Each overfitted candidate is cross-tested on all other universes and ranked by accuracy. Those that fail to generalize past a certain threshold are eliminated from the gene pool as overly specialized, fragile, or spurious.

3. *The true law survives.* The expression that consistently performs well across every universe captures the underlying mechanism, either as a singular survivor or a parsimonious ensemble of robust sub-laws. Even if a champion candidate is not the top performer in any single universe, its cross-universe consistency is what takes precedence.

4. *Asymptotic convergence dominates.* As the number of universes grows, the probability that a spurious formula survives decays exponentially. Conversely, assuming it lies within the grammar, the true law remains stable. Repeated tournament cycles therefore converge to the ground-truth equation.

In summary, CU-SRT offers a new mode of discovery: derivation by tournament. Rather than proving a law from first principles, the algorithm lets it emerge through cross-universe selection pressure. Overfitted candidates compete for universality in a global arena, and only the fittest survive. The result is a computational analog of theoretical discovery, where structure is not assumed but earned.

This evolutionary intuition dovetails with a long-standing scientific challenge. Throughout history, researchers have encountered messy data that were later found to obey simple, low-dimensional laws. Successes such as Kepler's third law and Newton's law of cooling are proof that elegant equations can be born from noisy measurements. However, in many modern domains, three obstacles remain:

(a) **Analytical Intractability:** First-principles derivation may be impossible because the governing dynamics are unknown, nonlinear, or prohibitively complex.

(b) **Heterogeneous Evidence:** Multiple independent data universes exist, each contaminated by its own noise and idiosyncrasies, making it impossible to get a single clean table of observations.

(c) **Overfitting Risk:** Contemporary symbolic regression engines can recover exact functional forms from data, but when trained on a single universe they more often than not end up with spurious artifacts that collapse out of sample.

1.2 Our Proposal

CU-SRT tackles these three obstacles simultaneously. We deliberately overfit symbolic regression models within each universe (addressing heterogeneity), then apply cross-universe generalization pressure (mitigating overfitting) and reward parsimony via a Minimum-Description -Length penalty (containing complexity). Only formulas that remain accurate and concise across all universes are allowed to survive. This way, we sidestep analytical intractability by letting the data do the heavy lifting of derivation.

By orchestrating mutation and selection across the universe set $\mathcal{U}$, CU-SRT turns empirical variety into a computational crucible, forging Dawkinsian replicator candidates into interpretable laws in the absence of classical derivation.

Conceived independently, CU-SRT nevertheless occupies the same problem space as earlier multi-system symbolic-discovery efforts. However, our framework diverges by casting cross-universe survival as an explicit tournament. CU-SRT yields formal generalization guarantees and domain-agnostic scalability by deliberately overfitting separate models in each universe and then culling those that fall short of cross-domain validity. What results is an adversarial mutation-and-selection loop that complements, rather than competes with, joint-invariance objectives.

2 Problem Statement

Let $\mathcal{U} = \{u_1, \dots, u_N\}$ be a family of universes, each providing a paired data set

$$\big(X^{(u)}, Y^{(u)}\big) \;=\; \big\{(x_t^{(u)}, y_t^{(u)}) \;\big|\; t = 1, \dots, T_u\big\}, \qquad u \in \mathcal{U},$$

where $x_t^{(u)} \in \mathbb{R}^d$ denotes the feature vector for observation t in universe u and $y_t^{(u)} \in \mathbb{R}$ (or $\mathbb{R}^m$) is the corresponding target (ground-truth response to be predicted).

Goal: Discover an analytic mapping $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^m$ that satisfies two competing criteria:

1. **Predictive fidelity** across all universes:

$$Y^{(u)} \;\approx\; \mathcal{L}\big(X^{(u)}\big), \qquad \forall\, u \in \mathcal{U}. \tag{1}$$

2. **Parsimony**: $\mathcal{L}$ should be as concise as possible to maximize interpretability and minimize over-parameterization.

2.1 Formal Optimization

Define the universe-specific generalization score

$$G_u(\varphi) \;=\; \max\Bigl\{0,\, 1 - \frac{\bigl\| Y^{(u)} - \varphi\bigl(X^{(u)}\bigr) \bigr\|_2^2}{\bigl\| Y^{(u)} \bigr\|_2^2}\Bigr\}, \tag{2}$$

which normalizes mean-squared error (MSE) so that $G_u = 1$ for a perfect fit and $G_u = 0$ for any fit no better than predicting zero. Thus every score satisfies $0 \leq G_u \leq 1$. We adopt hard clipping purely for algebraic clarity in the concentration bounds. Any bounded, monotone squash (e.g. logistic or exponential) would serve equally well, and sharper variants are left to future empirical work.

Let the cross-universe accuracy be the arithmetic mean

$$\bar{G}(\varphi) \;=\; \frac{1}{N} \sum_{u \in \mathcal{U}} G_u(\varphi).$$

(A weighted variant for heteroskedastic universes is given later in Section 5.1.)

Denote by $\ell(\varphi)$ the symbolic length (or any description-length proxy) of φ. The discovery problem is then cast as a multi-objective optimization:

$$\boxed{\mathcal{L}^\star \;=\; \arg\max_{\varphi \in \mathcal{F}} \Bigl\{\bar{G}(\varphi) \;-\; \lambda\, \ell(\varphi)\Bigr\}} \tag{3}$$

The Master Equation of CU-SRT: Law of Cross-Universe Survival

where $\lambda > 0$ trades off accuracy against complexity, and $\mathcal{F}$ is the finite hypothesis space of candidate expressions generated by the symbolic grammar $\mathcal{G}_{k,D}$ (i.e. all parse trees of depth $\leq D$ built from k primitives).[1]

2.2 Interpretation

Equation (3) distills Darwinian selection into algebra. It is natural selection as an argmax functional: maximize fitness, pay for complexity. A candidate φ that overfits its home universe yet fails elsewhere earns a low cross-universe score $\bar{G}(\varphi)$ and is eliminated. A highly accurate but baroque expression is handicapped by the complexity penalty $\lambda\ell(\varphi)$. In evolutionary terms, $\bar{G}$ measures an organism's fitness across environments, while $\lambda\ell(\varphi)$ represents the metabolic cost of carrying excess genetic baggage. The tournament therefore favors "universal minimalists" or formulas that achieve maximal transferability with minimal symbolic DNA. Only those that strike this balance survive as credible approximations of the hidden law $\mathcal{L}$.

[1] The MDL principle views each $\varphi \in \mathcal{F}$ as a codeword whose length $\ell(\varphi)$ supplies an a priori cost. Minimizing $\bar{G}(\varphi) - \lambda\ell(\varphi)$ therefore balances predictive accuracy against description length in an information-theoretic sense.

3 The CU-SRT Algorithm

CU-SRT proceeds in three phases: local mutation, cross-universe selection, and output synthesis. Algorithmic hyper-parameters are γ (minimum in-sample accuracy), λ (complexity penalty), and τ (survival threshold).

3.1 Phase A — Local Mutation

Overfitting births a riot of fragile genes,
from which evolution will later carve out what endures.

1. **Symbolic Search:** For every universe $u \in \mathcal{U}$ run a symbolic regression engine to obtain a ranked list of candidate formulas $\mathcal{C}_u = \{\varphi_{u,1}, \dots, \varphi_{u,M}\}$. Enforce a tight in-sample criterion

   $$G_u(\varphi_{u,j}) > \gamma, \qquad \text{where } G_u \text{ is defined in (2).}$$

   The gate γ is not a complexity tax but rather a heartbeat check: a formula must show the faintest pulse of explanatory power in its own habitat before joining the global gene pool. Over-elaborate or overfitted mutants are welcome for now. The harsher, complexity-aware selection pressure arrives in Phase B's cross-universe arena.[2] [3]

2. **Candidate Pool Aggregation:** Collect all local champions into the global mutation pool $\mathcal{C} = \bigcup_{u \in \mathcal{U}} \mathcal{C}_u$.

3.2 Phase B — Cross-Universe Selection

Cross-testing plays the role of natural selection.

1. **Generalization Score**[4]**:** For each $\varphi \in \mathcal{C}$, compute its average out-of-sample performance

   $$\bar{G}(\varphi) \;=\; \frac{1}{N} \sum_{v \in \mathcal{U}} G_v(\varphi).$$

2. **Tournament Score:** Combine accuracy and parsimony via

   $$\mathcal{T}(\varphi) \;=\; \bar{G}(\varphi) \;-\; \lambda\, \ell(\varphi), \tag{4}$$

   where $\ell(\varphi)$ is the symbol count of φ. A candidate now faces global selection: it must generalize across every universe but also pay a "metabolic" tax for each symbol it carries (strength minus load). Since $\sum_{\varphi \in \mathcal{F}} 2^{-\ell(\varphi)} \leq 1$ (Kraft inequality), the term $2^{-\ell(\varphi)}$ constitutes a valid prefix-free prior. Consequently, the additive penalty $\lambda\, \ell(\varphi)$ is an MDL code length that enforces parsimony without biasing the score scale.[5]

3. **Elimination:** Discard any candidate with $\mathcal{T}(\varphi) < \tau$. The surviving set is denoted $\mathcal{S}$.

[2] Optional grammar annealing reweighs primitives each generation; see Section 5.2.

[3] Optional causal-graph pruning removes sign-violators before Phase B; see Section 5.3.

[4] Optional weighted scoring for heteroskedastic universes; see Section 5.1.

[5] Optional Bayesian alternative replaces $\bar{G}(\varphi)$ with a sum of log–marginal likelihoods; see Section 5.4.

3.3 Phase C — Output Synthesis

The strongest lineage lives on.
Either a single champion reigns alone, or the survivors fuse into an ensemble.

- **Single Champion Minimum Description Length (MDL):** Select $\mathcal{L}^{\star} = \arg\max_{\varphi \in \mathcal{S}} \mathcal{T}(\varphi)$, yielding a unique closed-form law.

- **Tournament-Weighted Ensemble:** Produce a composite law

$$\mathcal{L}(x) \;=\; \frac{\sum_{\varphi \in \mathcal{S}} \mathcal{T}(\varphi)\, \varphi(x)}{\sum_{\varphi \in \mathcal{S}} \mathcal{T}(\varphi)},$$

which averages survivors in proportion to their tournament strength. When different survivors excel in distinct regions of the input space, their weighted blend could outperform any single law. Either the single strongest law rules alone, or the survivors join forces in an ensemble: a Darwinian republic chosen by $\mathcal{T}$. Therefore, we keep the single champion whenever its tournament lead is clear. Otherwise, we adopt an ensemble only if its out-of-sample score beats the champion by a small user-set margin.

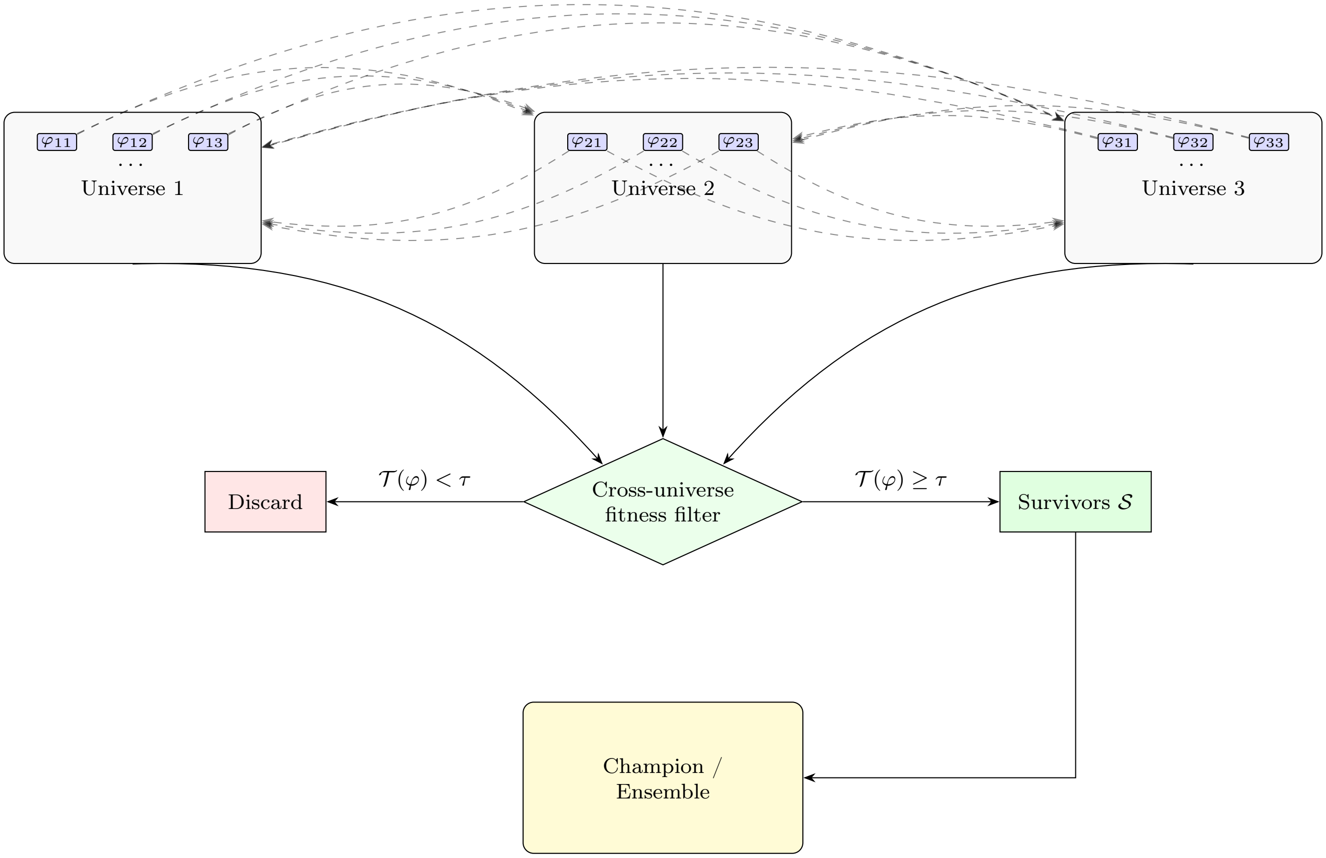

Figure 1: Cross-testing schematic. Each universe spawns several overfitted formulas (three shown). Every formula is evaluated on the other universes (dashed arrows), yielding its tournament score $\mathcal{T}(\varphi)$. Candidates with $\mathcal{T}(\varphi) < \tau$ are discarded; those with $\mathcal{T}(\varphi) \geq \tau$ form the survivor set $\mathcal{S}$ that feeds the champion/ensemble stage.

Lemma 1 (Ensemble Risk Upper Bound)**.** Let $\mathcal{S}$ be the survivor set and assign each $\varphi \in \mathcal{S}$ the non-negative weight $w_\varphi = \mathcal{T}(\varphi)$ with total weight $W := \sum_{\psi \in \mathcal{S}} w_\psi > 0$ and normalized weights $p_\varphi := w_\varphi / W$. Define the weighted ensemble

$$\mathcal{L}_{\text{ens}} \;=\; \sum_{\varphi \in \mathcal{S}} p_\varphi \, \varphi.$$

Assume the population mean-squared error of every survivor is bounded,

$$R_u(\varphi) := \mathbb{E}\big[(Y^{(u)} - \varphi(X^{(u)}))^2\big] \le \nu^2, \qquad \forall\, u,\ \varphi \in \mathcal{S}.$$

Then, for every universe u,

$$R_u\big(\mathcal{L}_{\text{ens}}\big) \;\le\; R_u\big(\mathcal{L}^\star\big) \;+\; \nu^2,$$

where $\mathcal{L}^\star := \arg\min_{\varphi \in \mathcal{S}} R_u(\varphi)$.

Proof. Because the squared-error loss $\ell(z) = \|z\|_2^2$ (not to be confused with the length penalty $\ell(\varphi)$) is convex, Jensen's inequality yields

$$R_u\big(\mathcal{L}_{\text{ens}}\big) \;=\; \mathbb{E}\Big[\ell\Big(\sum_\varphi p_\varphi \{Y^{(u)} - \varphi(X^{(u)})\}\Big)\Big] \;\le\; \sum_{\varphi \in \mathcal{S}} p_\varphi R_u(\varphi).$$

Splitting this sum at the champion $\mathcal{L}^\star$ gives

$$\sum_\varphi p_\varphi R_u(\varphi) \;=\; R_u(\mathcal{L}^\star) + \sum_\varphi p_\varphi \big[R_u(\varphi) - R_u(\mathcal{L}^\star)\big] \;\le\; R_u(\mathcal{L}^\star) + \sum_\varphi p_\varphi \, \nu^2 \;=\; R_u(\mathcal{L}^\star) + \nu^2,$$

because $\sum_\varphi p_\varphi = 1$. □

Lemma 1 guarantees the ensemble can inflate risk by at most ν^2, while Example 1 (below) shows it would most likely reduce risk due to variance cancellation.

Example 1 (Variance cancellation in a two-law pool)**.** Let $S = \{\varphi_1, \varphi_2\}$ and suppose each survivor predicts $\varphi_i\big(X^{(u)}\big) \;=\; \mathcal{L}\big(X^{(u)}\big) + \eta_i$, where the noises $\eta_i \sim \mathcal{N}\big(0, \nu^2/2\big)$ are independent of each other and of $Y^{(u)} = \mathcal{L}\big(X^{(u)}\big) + \eta$ with $\eta \sim \mathcal{N}\big(0, \nu^2/2\big)$. Then every survivor satisfies $R_u(\varphi_i) = \mathbb{E}\big[(\eta - \eta_i)^2\big] = \nu^2$, meeting the bound in Lemma 1. The equal-weight ensemble $\mathcal{L}_{\text{ens}} = \frac{1}{2}\big(\varphi_1 + \varphi_2\big)$ incurs error $\eta - \frac{1}{2}\big(\eta_1 + \eta_2\big)$ and hence risk $R_u\big(\mathcal{L}_{\text{ens}}\big) = \frac{\nu^2}{4} + \frac{\nu^2}{2} = \frac{3}{4}\nu^2$, which is strictly smaller than the risk of either survivor. Thus Lemma 1 allows the ensemble to outperform every individual law whenever their errors are not perfectly correlated.

Proposition 1 (Risk scaling with n independent survivors)**.** *Let $S = \{\varphi_1, \ldots, \varphi_n\}$ and assume each survivor predicts $\varphi_i\big(X^{(u)}\big) \;=\; \mathcal{L}\big(X^{(u)}\big) + \eta_i$, where the centered noises $\eta_i \overset{i.i.d.}{\sim} \mathcal{N}\big(0, \nu^2/2\big)$ are independent of one another and of $Y^{(u)} = \mathcal{L}\big(X^{(u)}\big) + \eta$ with $\eta \sim \mathcal{N}\big(0, \nu^2/2\big)$. Then every survivor satisfies $R_u(\varphi_i) = \nu^2$, meeting the assumption of Lemma 1. Using equal weights $p_\varphi = 1/n$, the ensemble $\mathcal{L}_{\text{ens}} = \frac{1}{n}\sum_{i=1}^n \varphi_i$ has population risk*

$$R_u\big(\mathcal{L}_{\text{ens}}\big) = \frac{n+1}{2n}\,\nu^2\ ,$$

which is strictly smaller than ν^2 for every $n \ge 2$, and it tends to $\nu^2/2$ as $n \to \infty$.

Proof. Since η is independent of the η_i,

$$R_u\big(\mathcal{L}_{\text{ens}}\big) = \operatorname{Var}\Big(\eta - \frac{1}{n}\sum_{i=1}^{n}\eta_i\Big) = \frac{\nu^2}{2} + \frac{\nu^2}{2n} = \frac{n+1}{2n}\,\nu^2.$$

□

CU-SRT Workflow

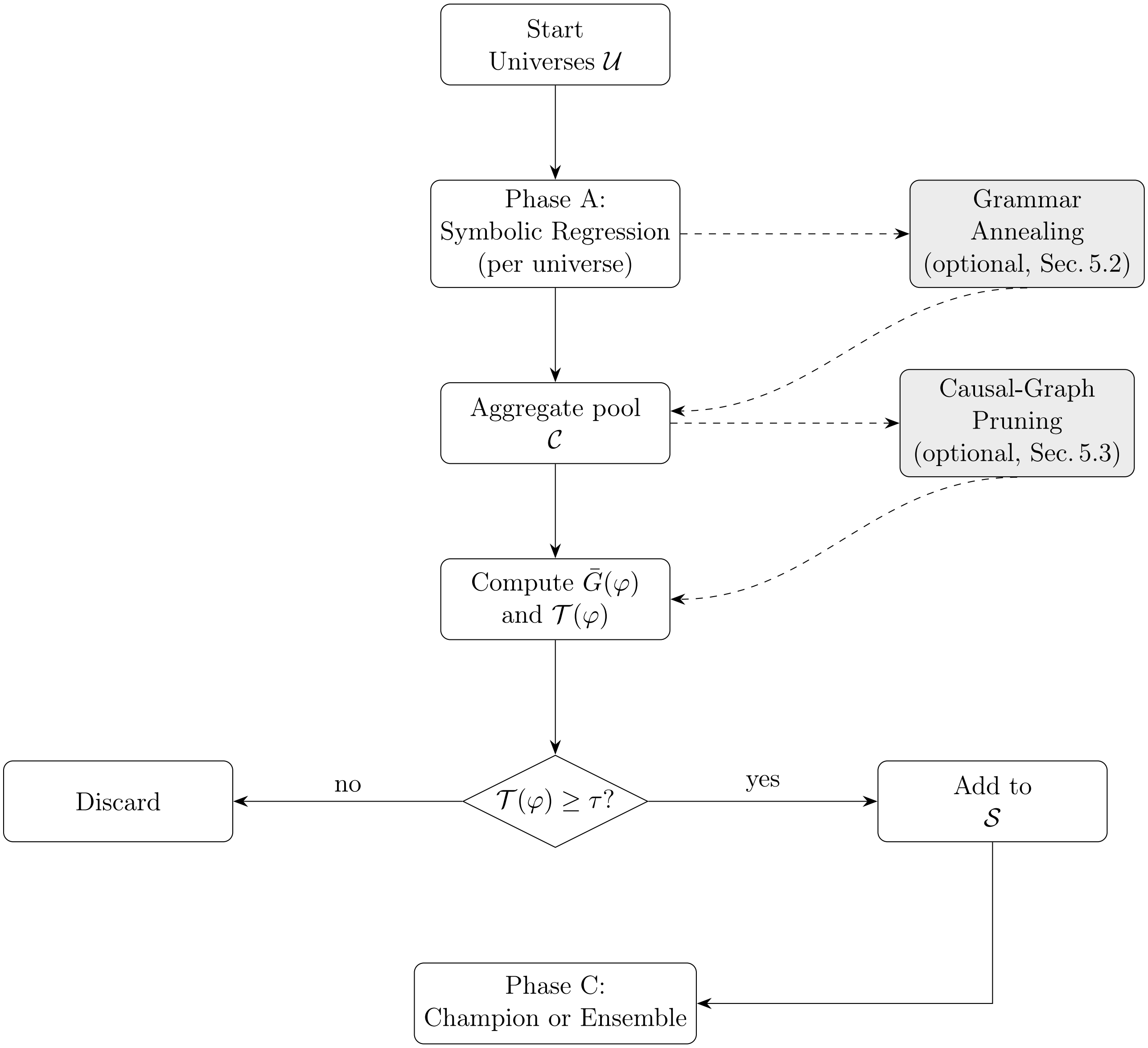

Figure 2: CU-SRT workflow. Solid arrows show the default pipeline (Phase A → pool → scoring → selection). Dashed arrows indicate two independent optional layers: stochastic grammar annealing (Section 5.2) applied directly after local mutation and causal-graph pruning (Section 5.3) applied just before cross-universe scoring.

4 Theoretical Guarantees

The CU-SRT is based on 3 main principles: overfitting as mutation, cross-universe survival, and complexity bias. Together they form an evolutionary analog of scientific induction, with formal guarantees under mild assumptions.

4.1 Overfitting as Mutation

Symbolic regression applied to a single universe u searches an exponentially large space of parse trees.[6] When the in-sample threshold $G_u(\varphi) > \gamma$ is set close to 1, the resulting mutation pool $\mathcal{C}_u = \{\varphi_{u,1}, \ldots, \varphi_{u,M}\}$ is guaranteed (in the limit of exhaustive or unbiased SR search) to contain:

1. The true law $\mathcal{L}$ restricted to universe u (if expressible within the grammar).

2. Numerous parasite formulas that incorporate idiosyncratic noise.

Thus, local overfitting serves the same role as genetic mutation. It creates high-variance diversity around the neighborhood of $\mathcal{L}$.

4.2 Cross-Universe Survival

Assume every universe $u \in \mathcal{U}$ is sampled from the same data-generating mechanism $Y = \mathcal{L}(X) + \eta$, where the noise η has zero mean and finite variance. For any spurious formula $\tilde{\varphi}$ that deviates from $\mathcal{L}$ by at least $\Delta > 0$ in expectation, the cross-universe score

$$\bar{G}(\tilde{\varphi}) = \frac{1}{N} \sum_{u \in \mathcal{U}} G_u(\tilde{\varphi})$$

converges to a value strictly less than the score of $\mathcal{L}$ as $N \to \infty$. Because each per-universe accuracy score is clipped to the unit interval, $0 \leq G_u(\varphi) \leq 1$, Hoeffding's inequality with range 1 applies. Let $\zeta > 0$ be an arbitrary tolerance parameter. A Chernoff bound yields

$$\Pr\big\{\bar{G}(\tilde{\varphi}) \geq \bar{G}(\mathcal{L}) - \zeta\big\} \ \leq \ \exp\big(-2N\zeta^2\big), \tag{5}$$

so the likelihood of a non-truthful formula surviving the tournament decays exponentially with the number of universes.

4.3 Complexity Bias

To break ties among formulas with similar $\bar{G}$ values we subtract a length penalty $\lambda\ell(\varphi)$ (Equation (4)). This acts as an information-theoretic prior in the spirit of Minimum Description Length (MDL): shorter programs are deemed a priori more plausible. Let $\mathcal{S}_L = \{\varphi \in \mathcal{S} : \ell(\varphi) = L\}$ be the survivor subset of length L. Under a Kraft–McMillan coding interpretation, the weighted score $2^{-\ell(\varphi)}$ is proportional to the inverse code-length probability, ensuring the selection process remains prefix-free and self-consistent.

[6] For a grammar with k primitive functions and maximum depth D, the search space is $\mathcal{O}(k^D)$.

4.4 Asymptotic Optimality

Combining the Chernoff bound in Equation (5) with a mild length-minimality assumption (stated next) yields the finite-sample guarantee in the corollary below.

Assumption 1 (Grammar-expressivity)**.** The symbolic grammar $\mathcal{G}_{k,D}$ contains at most k primitives (functions or terminals) and allows parse-trees of depth at most D. Consequently its cardinality is bounded by $|\mathcal{G}_{k,D}| \leq k^D =: C_{k,D}$, and we assume the true law $\mathcal{L}$ lies in $\mathcal{G}_{k,D}$.

Assumption 2 (Length-minimality)**.** No spurious formula is shorter than the true law: $\ell(\tilde{\varphi}) \geq \ell(\mathcal{L})$ for every $\tilde{\varphi} \neq \mathcal{L}$.

Proposition 2 (Asymptotic consistency under noise)**.** *Let $\mathcal{L}$ be the shortest formula in $\mathcal{F}$ that minimizes the population risk and denote its cross-universe generalization score by $\bar{G}(\mathcal{L})$. Assume 1–2 hold and that there exists a positive margin*

$$\Delta \;=\; \bar{G}(\mathcal{L}) \;-\; \sup_{\tilde{\varphi}\neq\mathcal{L}} \bar{G}(\tilde{\varphi}) \;>\; 0.$$

Then for every spurious $\tilde{\varphi} \neq \mathcal{L}$

$$\Pr\{\mathcal{T}(\tilde{\varphi}) \geq \mathcal{T}(\mathcal{L})\} \;\leq\; \exp\bigl(-2N\Delta^2\bigr).$$

Consequently, letting the number of universes $N \to \infty$ (with fixed $\lambda > 0$) drives the mis-selection probability to zero. If, in addition, every spurious formula is strictly longer than $\mathcal{L}$ (so $\ell(\tilde{\varphi}) > \ell(\mathcal{L})$), then letting $\lambda \to \infty$ (with fixed $N > 0$) also drives the mis-selection probability to zero. In either limit the CU-SRT tournament selects $\mathcal{L}$ with probability 1.

Proof. Write the tournament score as $\mathcal{T}(\varphi) = \bar{G}(\varphi) - \lambda\ell(\varphi)$. For any spurious $\tilde{\varphi} \neq \mathcal{L}$ we have

$$\bigl\{\mathcal{T}(\tilde{\varphi}) \geq \mathcal{T}(\mathcal{L})\bigr\} \iff \Bigl\{\bar{G}(\tilde{\varphi}) \geq \bar{G}(\mathcal{L}) - \lambda\bigl[\ell(\mathcal{L}) - \ell(\tilde{\varphi})\bigr]\Bigr\}. \tag{6}$$

If $\ell(\tilde{\varphi}) > \ell(\mathcal{L})$, the right-hand threshold in (6) exceeds $\bar{G}(\mathcal{L})$, while the margin assumption $\bar{G}(\tilde{\varphi}) \leq \bar{G}(\mathcal{L}) - \Delta$ still holds; hence the event has probability at most

$$\exp\Bigl(-2N\bigl[\Delta + \lambda\bigl(\ell(\tilde{\varphi}) - \ell(\mathcal{L})\bigr)\bigr]^2\Bigr),$$

which vanishes as $N \to \infty$ or $\lambda \to \infty$. Thus only the case $\ell(\tilde{\varphi}) = \ell(\mathcal{L})$ requires a deviation bound.

Chernoff Bound with $\zeta = \Delta$: With equal lengths the condition (6) reduces to $\bar{G}(\tilde{\varphi}) \geq \bar{G}(\mathcal{L})$. Applying Hoeffding's inequality to the i.i.d. universe scores gives

$$\Pr\bigl[\bar{G}(\tilde{\varphi}) \geq \bar{G}(\mathcal{L})\bigr] \;\leq\; \exp\bigl(-2N\Delta^2\bigr). \tag{7}$$

where $\Delta = \bar{G}(\mathcal{L}) - \sup_{\psi\neq\mathcal{L}} \bar{G}(\psi) > 0$.

Since the probability is exponentially small for $\ell(\tilde{\varphi}) > \ell(\mathcal{L})$ and bounded by (7) when $\ell(\tilde{\varphi}) = \ell(\mathcal{L})$, we obtain

$$\Pr\bigl[\mathcal{T}(\tilde{\varphi}) \geq \mathcal{T}(\mathcal{L})\bigr] \;\leq\; \exp\bigl(-2N\Delta^2\bigr).$$

Hence letting $N \to \infty$ drives the mis-selection probability to 0. If every spurious formula is in fact strictly longer than $\mathcal{L}$, the event becomes negligible once $\lambda \to \infty$ with fixed N, completing the proof. □

Corollary 1 (Finite sample guarantee)**.** Let

$$\Delta \;=\; \bar{G}(\mathcal{L}) \;-\; \max_{\varphi \neq \mathcal{L}} \bar{G}(\varphi) \;>\; 0$$

be the population generalization gap, and fix a confidence level $0 < \beta < 1$. Here $|\mathcal{C}|$ denotes the total number of candidate formulas under consideration. For example, the tournament pool, bounded above by $C_{k,D}$ from Assumption 1. If the number of universes satisfies

$$N \;\geq\; \frac{\log|\mathcal{C}| + \log(1/\beta)}{2\Delta^2},$$

then the CU-SRT tournament selects the true law $\mathcal{L}$ with probability at least $1 - \beta$.

Proof. For every spurious candidate $\tilde{\varphi} \neq \mathcal{L}$ the Chernoff bound (5) gives $\Pr\big[\bar{G}(\tilde{\varphi}) \geq \bar{G}(\mathcal{L}) - \Delta\big] \leq \exp(-2N\Delta^2)$. Applying a union bound over the finite pool $\mathcal{C}$ yields

$$\Pr\Big[\exists\,\tilde{\varphi} \neq \mathcal{L} : \bar{G}(\tilde{\varphi}) \geq \bar{G}(\mathcal{L}) - \Delta\Big] \;\leq\; |\mathcal{C}|\,\exp(-2N\Delta^2).$$

Requiring the right-hand side to be no larger than β and solving for N gives $N \geq (\log|\mathcal{C}| + \log(1/\beta))/(2\Delta^2)$, as claimed. □

Remark 1 (Effect of pool size and universes)**.** *Increasing the mutation pool* $|\mathcal{C}|$ *(e.g. by lowering the in-sample threshold* γ *or running the symbolic regression longer) raises the probability that the true law* $\mathcal{L}$ *lies in* $\mathcal{C}$*, while increasing the number of universes* N *exponentially decreases the probability that any spurious* $\tilde{\varphi} \neq \mathcal{L}$ *survives (Chernoff bound* (5)*). Hence larger* $(|\mathcal{C}|, N)$ *jointly tighten the tournament's convergence to* $\mathcal{L}$*.*

4.5 Adaptive Survival Thresholds and Early Termination

The baseline CU-SRT uses a fixed survival threshold τ. Any candidate with $\mathcal{T}(\varphi) < \tau$ is culled immediately (as defined in Phase B, Step 3). While a static bar makes sense when all universes are known right off the bat, it can be sub-optimal or computationally wasteful when data arrives incrementally or when the number of universes N is small. We therefore introduce a more realistic option: an adaptive threshold τ_t that tightens over successive tournament rounds t, prunes weak formulas early, and cuts symbolic regression work to $\mathcal{O}\big(\log|\text{pool}|\big)$ while the MDL penalty $\lambda\ell(\varphi)$ remains unchanged.

Dynamic τ Schedule: Let $\mathcal{C}^{(t)}$ denote the candidate pool at round t and let

$$\mu_t \;=\; \frac{1}{|\mathcal{C}^{(t)}|}\sum_{\varphi \in \mathcal{C}^{(t)}} \mathcal{T}_t(\varphi), \qquad \rho_t^2 \;=\; \frac{1}{|\mathcal{C}^{(t)}|}\sum_{\varphi \in \mathcal{C}^{(t)}} \big(\mathcal{T}_t(\varphi) - \mu_t\big)^2,$$

where $\rho_t := \sqrt{\rho_t^2}$ is the standard deviation of $\mathcal{T}_t(\varphi)$ at round t.

Choose parameters $\alpha \in (0,1)$ (aggression: lower α implies harsher cull, measured in standard-deviation units) and $q \in (0,1)$ (minimum discard fraction). We set

$$\tau_t \;=\; \max\Big\{\mu_t - \alpha\rho_t,\; \text{Quantile}_q\big(\{\mathcal{T}_t(\varphi)\}\big)\Big\}, \tag{8}$$

where Quantile_q keeps at most the top $(1-q)$ fraction of candidates. Thus τ_t rises automatically as the score distribution converges, emulating increasing selection pressure.

Algorithmic Update: Phase B now loops over rounds $t = 1, 2, \ldots$:

1. *Adaptive Elimination:* Discard candidates with $\mathcal{T}_t(\varphi) < \tau_t$. If $\tau_t = \tau_{t-1}$ and the survivor set $\mathcal{S}^{(t)} = \mathcal{S}^{(t-1)}$, **terminate** (early-stopping criterion).

2. *Optional Data Refresh:* If new universes have arrived, recompute $\bar{G}_{t+1}(\varphi)$ and $\mathcal{T}_{t+1}(\varphi)$ for all surviving φ and continue.

Practical Note: Because each newly arriving universe immediately re-computes μ_t, ρ_t, τ_t for all candidates, we keep a single shrinking survivor set $S(t)$ instead of archiving separate pools for every past τ. This avoids duplicate scoring passes while still allowing late universes to eject earlier survivors when warranted.

Consistency and Compute Savings:

Proposition 3 (Finite-round consistency)**.** *Assume the conditions of Proposition 2 hold and suppose rounds continue until the early-stopping test above is triggered at round T. Then the champion $\mathcal{L}_T^\star = \arg\max_{\varphi \in \mathcal{S}^{(T)}} \mathcal{T}_T(\varphi)$ satisfies*

$$\Pr\Big\{\mathcal{T}_T(\mathcal{L}_T^\star) < \mathcal{T}(\mathcal{L}) - \zeta\Big\} \;\le\; |\mathcal{C}^{(1)}| \, \exp\!\big(-2\, N_T\, \zeta^2\big),$$

where N_T is the number of universes evaluated up to round T. In particular, if

$$N_T \;\ge\; \left\lceil \tfrac{\log|\mathcal{C}^{(1)}| + \log(1/\beta)}{2\,\zeta^2} \right\rceil,$$

the dynamic-τ algorithm selects a law within ζ of the optimal tournament score $\mathcal{T}(\mathcal{L})$ with probability at least $1 - \beta$.

Sketch. At each round the adaptive rule (8) removes at least the lowest-scoring q fraction of candidates. Hence after T rounds the survivor pool is at most $|\mathcal{C}^{(1)}|(1-q)^T$. Apply the Chernoff bound used in Equation (5) and union-bound over the shrinking pool. The details mirror the proof of Corollary 1 but with N replaced by N_T. □

Corollary 2 (Geometric contraction of the candidate pool)**.** Let $|\mathcal{C}^{(1)}|$ be the initial pool size and let $q \in (0,1)$ be the minimum survivor fraction in (8). After t adaptive rounds

$$|\mathcal{C}^{(t)}| \;\le\; |\mathcal{C}^{(1)}| \, (1-q)^t,$$

so the pool size decays geometrically. In particular,

$$|\mathcal{C}^{(t)}| < 1 \quad \text{whenever} \quad t \;>\; \frac{\log|\mathcal{C}^{(1)}|}{\log\!\big(\frac{1}{1-q}\big)} \;=\; \log_{1/(1-q)} |\mathcal{C}^{(1)}|.$$

Proof. Rule (8) discards at least the lowest-scoring q fraction each round, so $|\mathcal{C}^{(t)}| \le (1-q)\,|\mathcal{C}^{(t-1)}|$. Induction yields the geometric bound, and the stopping index follows by solving $(1-q)^t < 1/|\mathcal{C}^{(1)}|$ for t. □

Total Candidate Evaluations: Summing the geometric bound above gives

$$\textstyle\sum_{t \ge 1} |\mathcal{C}^{(t)}| \;\le\; |\mathcal{C}^{(1)}|/q,$$

so adaptive pruning costs at most a $1/q$-fold pass over the initial pool.

5 Optional Enhancements and Extensions

5.1 Universe-Weighted Scores

Data-rich universes get louder megaphones.

Motivation: In some domains certain universes are noticeably noisier or smaller. It would be beneficial in this case to assign weights to each one, so that the aggregate score is proportional to how valuable its input actually is. Define empirical noise variance

$$\hat{\sigma}_u^2 := \frac{1}{T_u}\sum_{t=1}^{T_u}\left(y_t^{(u)} - \bar{y}^{(u)}\right)^2, \qquad \bar{y}^{(u)} := \frac{1}{T_u}\sum_t y_t^{(u)},$$

and set inverse-variance weights

$$w_u \ := \ \frac{T_u/\hat{\sigma}_u^2}{\sum_{v\in\mathcal{U}} T_v/\hat{\sigma}_v^2}, \qquad \sum_{u\in\mathcal{U}} w_u = 1.$$

Replace the uniform mean by

$$\bar{G}_w(\varphi) \ = \ \sum_{u\in\mathcal{U}} w_u\, G_u(\varphi),$$

and use $\bar{G}_w(\varphi)$ in Equation (4). Because each $G_u(\varphi)\in[0,1]$ and $\sum_u w_u = 1$, the usual Hoeffding bound with range 1 still applies. (The fully weighted version is proved in Section 5.1.1). The factor T_u rewards universes with more data, while $1/\hat{\sigma}_u^2$ down-weights high-variance environments, mirroring weighted least squares under heteroskedastic noise. Computing $\hat{\sigma}_u^2$ adds $\mathcal{O}(T_u)$ time once per universe and does not affect SR-runtime complexity.

5.1.1 Hoeffding Bound with Universe Weights

Lemma 2 (Weighted Hoeffding)**.** Let $\{G_u(\varphi)\}_{u\in\mathcal{U}}$ be independent random variables bounded in $[0,1]$, let $w_u \geq 0$ satisfy $\sum_{u\in\mathcal{U}} w_u = 1$, and define

$$\bar{G}_w(\varphi) \ = \ \sum_{u\in\mathcal{U}} w_u\, G_u(\varphi).$$

For any deviation level $\epsilon > 0$,

$$\Pr\!\left[\bar{G}_w(\varphi) - \mathbb{E}\,\bar{G}_w(\varphi) \geq \epsilon\right] \ \leq \ \exp\!\left(-\tfrac{2\epsilon^2}{N\,(\max_u w_u)^2}\right),$$

and the same inequality holds for the lower tail. When $w_u = 1/N$ (uniform case), the exponent reduces to $-2N\epsilon^2$.

Sketch. Set $\tilde{G}_u := w_u G_u(\varphi) \in [0, w_u]$ and apply Hoeffding's inequality to $\sum_u(\tilde{G}_u - \mathbb{E}\,\tilde{G}_u)$. The range of each term is w_u, so

$$\Pr\left[\bar{G}_w - \mathbb{E}\,\bar{G}_w \geq \epsilon\right] \ \leq \ \exp\!\left(-\tfrac{2\epsilon^2}{\sum_u w_u^2}\right).$$

Because $\sum_u w_u^2 \ \leq \ \max_u w_u^2 \sum_u 1 \ = \ N \max_u w_u^2$, the stated bound follows. The lower-tail bound is analogous. □

Implications for Theoretical Guarantees (Section 4): Because each $G_u(\varphi)$ remains bounded in $[0, 1]$ and the weighted bound reduces to the uniform case when $w_u \leq 1/N$, all generalization and sample complexity results of Section 4 hold exactly the same after replacing the unweighted mean $\bar{G}$ with its weighted counterpart $\bar{G}_w$.

5.2 Stochastic Grammar Annealing

Promote the useful. Demote the idle. Never delete.

Motivation: A rich primitive set $\mathcal{F}_0$ $(+, -, \times, \div, \sin, \exp, \log, \dots)$ is crucial for early exploration at the local level. However, the same breadth inflates search complexity and encourages universe-specific overfitting on the global stage. Stochastic grammar annealing counters this by demoting the sampling probability of primitives that prove to be unhelpful across universes while never actually deleting any primitive outright.

Notation:

- $\mathcal{F}_0$ — initial function/operator set.
- $p_t(f)$ — probability of drawing primitive $f \in \mathcal{F}_0$ at tournament generation t.
- $\mathcal{C}_t$ — survivor pool after Phase B in generation t.
- T_t — annealing temperature; lower T_t means stronger bias toward frequently useful primitives.
- ε — exploration floor $(0 < \varepsilon \ll 1/|\mathcal{F}_0|)$; (not to be confused with $\epsilon > 0$ in Lemma 2).
- χ — re-heat trigger threshold (score-drop tolerance).
- $\Delta_t(f)$ — loss for primitive f at generation t $(1 - s_t(f))$; (not to be confused with the population margin Δ in Section 4).

Update Rule: After each Phase B:

(a) *Usage score*: for each primitive compute $s_t(f) = \dfrac{1}{|\mathcal{C}_t|} \sum_{\varphi \in \mathcal{C}_t} \mathbf{1}\{f \text{ appears in } \varphi\}$.

(b) *Loss*: $\Delta_t(f) = 1 - s_t(f)$ captures how seldom f helped.

(c) *Probability update*:

$$p_{t+1}(f) \;=\; \frac{\varepsilon + \exp\bigl(-\Delta_t(f)/T_t\bigr)}{\displaystyle\sum_{g \in \mathcal{F}_0} \Bigl[\varepsilon + \exp\bigl(-\Delta_t(g)/T_t\bigr)\Bigr]}.$$

Phase A of the next generation samples primitives using p_{t+1}, so under-performing operators become progressively rarer but never vanish.

Floor Property: [7] Because ε appears in every numerator, every primitive retains a non-zero chance of being drawn:

$$p_{t+1}(f) \;\geq\; \frac{\varepsilon}{|\mathcal{F}_0|\,\varepsilon + \sum_g \exp(-\Delta_t(g)/T_t)} \;>\; 0.$$

Thus, even the least-used operator always retains a non-zero chance of being sampled in future generations.

Temperature Schedule and Re-heating: A geometric decay $T_t = T_0\theta^t$ ($T_0 \approx 1,\ 0 < \theta < 1$) should theoretically provide a robust exploration–exploitation balance. If the cross-universe score $\bar{G}(\varphi^\star)$ drops by more than a user-defined χ for something like two consecutive generations (or another short patience window appropriate to the application), temporarily raise T_{t+1} back toward T_0 ("re-heat") to restore exploration.

Practical Notes: Begin annealing after the first complete Phase B, once usage counts are trustworthy. Keep $\varepsilon > 0$ so no primitive is ever removed. Cool T_t geometrically and re-heat if the champion's score drops sharply, avoiding lock-in. This sharpens the MDL bias, trims the grammar, accelerates search, and purges gimmicks.

5.3 Causal-Graph Pruning Layer

Formulas that defy the food chain are culled before the hunt begins.

Motivation: In many domains, there exists a baseline qualitative direction of influence between variables in order for things to make sense. For example: an increase in temperature cannot reduce ideal-gas pressure. Such sign-certainties can be encoded as a signed causal set

$$\mathcal{M} \;=\; \big\{(x_i, \xi_i)\big\}_{i=1}^{d}, \qquad \xi_i \in \{+1, -1\},$$

where ξ_i specifies that every admissible law must satisfy $\partial_{x_i}\varphi > 0$ (for $\xi_i = +1$) or $\partial_{x_i}\varphi < 0$ (for $\xi_i = -1$) wherever the derivative is well-defined.

Pruning Rule: Let $\mathcal{C}$ denote the aggregate pool produced after local mutation (Section 3). Draw a mini-batch $\{x^{(m)}\}_{m=1}^{B}$ from the combined data of all universes, where $B \in \mathbb{N}$ denotes the user-chosen mini-batch size (number of sample points at which first-order gradients are evaluated and not to be confused with Phase B). For each candidate $\varphi \in \mathcal{C}$ compute finite-difference gradients $\partial_{x_i}\varphi\big(x^{(m)}\big)$. The candidate is discarded if

$$\exists\,(x_i, \xi_i) \in \mathcal{M},\ \exists\, m \leq B\ :\ \xi_i\,\partial_{x_i}\varphi\big(x^{(m)}\big) \leq 0. \tag{9}$$

Otherwise φ proceeds unchanged to Phase B. To mitigate spurious rejections caused by numerical noise, introduce a tolerance $\kappa \in [0, 1)$ and prune only when the violation fraction exceeds κ. Choosing κ in the small but non-zero range (e.g. around 0.05-0.15) should provide a modest buffer against numerical fluctuations and still allow for timely pruning.

Where It Runs: Equation (9) is applied immediately before Phase B (Fig. 2). The filter removes sign-inconsistent formulas early, thereby reducing the cross-universe workload

[7] A small floor prevents extinction of potentially useful operators and guarantees continual exploration.

and ensuring that all survivors respect the supplied causal semantics. Because the rule acts solely on $\mathcal{C}$, the theoretical results of Section 4 and all Phase B scoring equations remain valid without modification. One merely replaces the original pool by its sign-consistent subset.

Practical Notes: Gradient screening adds $O\big(B \cdot |\mathcal{C}|\big)$ flops per round. With $B \ll T_u$ and vectorized finite differences, the extra time and memory are negligible. Apply the filter only to variables with indisputable sign constraints and leave uncertain links free. This way it should sharpen interpretability and prune the pool without altering CU-SRT's agnostic core.

5.4 Bayesian Tournament Scoring

Selection now measures likelihood. The genes most certain to endure lead the lineage.

Motivation: The default CU-SRT score treats each universe's hit-rate $G_u(\varphi)$ as a single point estimate and enforces parsimony only through the additive MDL term $\lambda\,\ell(\varphi)$. A Bayesian score uses the entire likelihood of the observed hits, thereby (i) weighting universes by their sample size, (ii) shrinking noisy accuracies toward $\frac{1}{2}$, and (iii) injecting an automatic Occam factor that penalizes over-flexible formulas even before the MDL term is applied.

Per-universe Evidence: For universe u let $n_u := T_u$ denote its sample size and $k_u(\varphi) := n_u G_u(\varphi)$ the number of "hits" (correct predictions) a candidate law φ achieves on those points. Since $k_u(\varphi)$ can be fractional, we use the continuous Beta–binomial extension, which coincides with the ordinary Beta–Binomial when k_u is an integer. With the non-informative Jeffreys prior Beta$\big(\frac{1}{2}, \frac{1}{2}\big)$ on the true hit-probability $p_u(\varphi)$, the marginal likelihood (Beta–Binomial evidence) is

$$\mathrm{ML}_u(\varphi) = \frac{\Gamma\big(k_u(\varphi) + \frac{1}{2}\big)\,\Gamma\big(n_u - k_u(\varphi) + \frac{1}{2}\big)}{\Gamma\big(k_u(\varphi) + 1\big)\,\Gamma\big(n_u - k_u(\varphi) + 1\big)\,\Gamma(\frac{1}{2})^2}.$$

(The uniform prior Beta$(1,1)$ is a one-line swap and yields the simpler but less discriminative factor $(n_u + 1)^{-1}$.)

Bayesian Tournament Score: Replace the accuracy term in Equation (4) by the sum of log-evidences:

$$\mathcal{T}_{\mathrm{Bayes}}(\varphi) = \sum_{u \in \mathcal{U}} \log\big[\mathrm{ML}_u(\varphi)\big] - \lambda\,\ell(\varphi).$$

Taking logs prevents underflow and converts the evidence product into a sum.

Where It Runs: In Phase B–Step 2 (Section 3) one can replace the default score by $\mathcal{T}_{\mathrm{Bayes}}$ without touching the elimination or champion-selection logic; no diagram changes are required.

- For large n_u, $\log \mathrm{ML}_u(\varphi) \approx k_u \log G_u + (n_u - k_u) \log(1 - G_u) - \frac{1}{2} \log n_u$, i.e. the Bayesian score is the usual log-likelihood plus an $O(\log n_u)$ correction.
- Because $\log \mathrm{ML}_u \leq 0$ and $|\log \mathrm{ML}_u| \leq n_u$, all generalization bounds in Section 4 stay valid when $\bar{G}$ is replaced by Bayesian evidence.
- Extra cost is one evaluation of $\log \Gamma(\cdot)$ per universe–law pair.

6 Conclusion

CU-SRT turns the search for scientific law into a type of computational Darwinism. Within each universe, symbolic regression over-expresses its genes, spawning a chaos of candidate equations. When those equations migrate across universes, they face an unforgiving ecology: only forms that remain predictive everywhere, and do so with as little symbolic DNA as possible, avoid extinction. Complexity penalties supply the metabolic cost that keeps bloat in check, while the exponential-decay guarantee ensures that as the number of universes grows, impostors are driven out of the pool. What remains is either a compact survivor or a collective Darwinian republic, each forged by the trials of cross-environment adversity into a generalized law. In this way, CU-SRT distills the classical dialectic of conjecture and refutation into an algorithmic tournament, converting empirical diversity from a nuisance into the very pressure that shapes explanatory form.

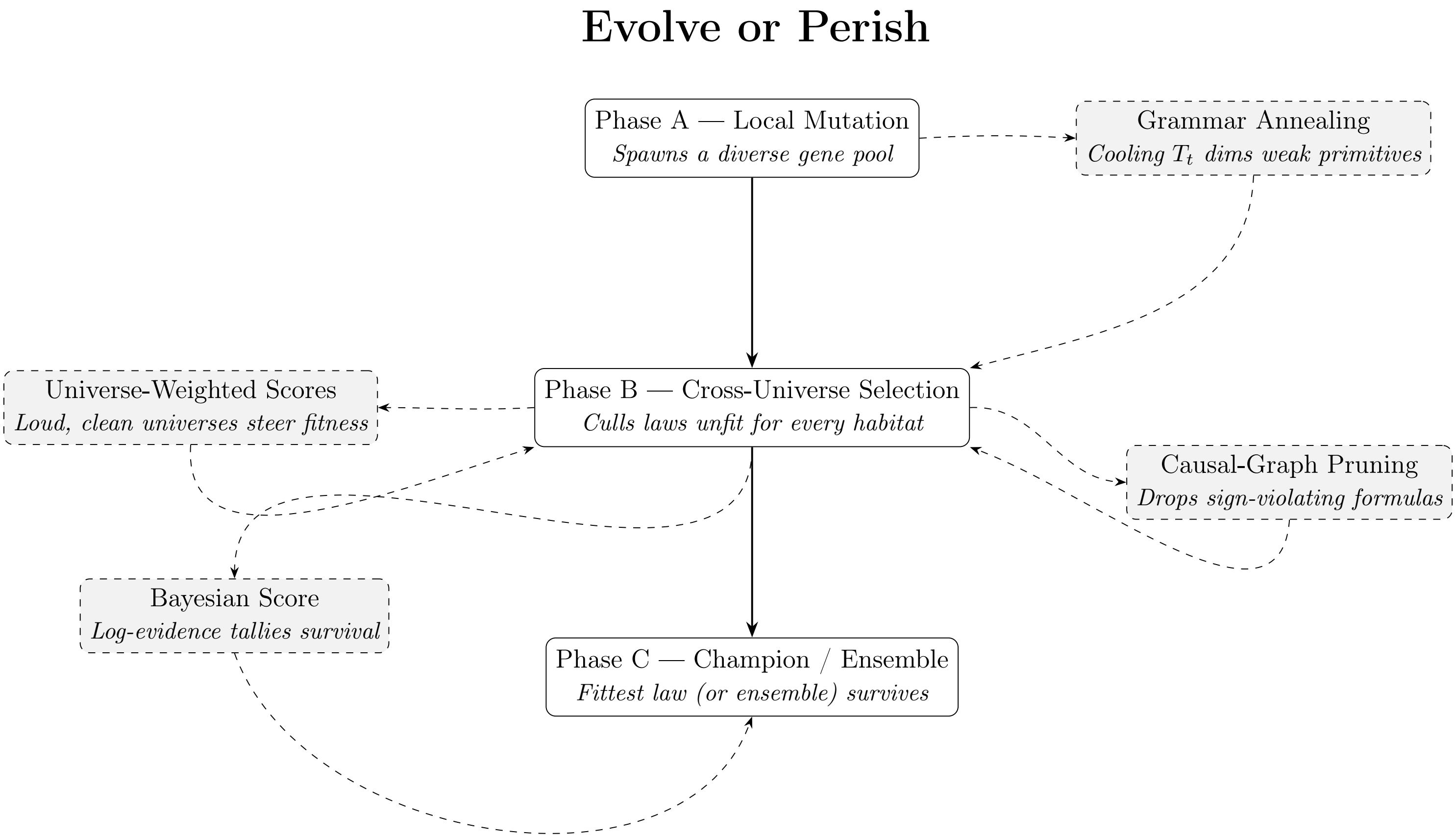
Evolve or Perish
Phase A — Local Mutation
Spawns a diverse gene pool
Grammar Annealing
Cooling T_t dims weak primitives
Universe-Weighted Scores
Loud, clean universes steer fitness
Phase B — Cross-Universe Selection
Culls laws unfit for every habitat
Causal-Graph Pruning
Drops sign-violating formulas
Bayesian Score
Log-evidence tallies survival
Phase C — Champion / Ensemble
Fittest law (or ensemble) survives

Figure 3: CU-SRT roadmap. Solid arrows and white boxes form the core pipeline (Phase A → Phase B → Phase C). Dashed arrows and gray dashed boxes mark optional layers. Each curves out from, and re-joins, the spine so any subset of options still yields a valid execution path. Box labels give a compact technical or evolutionary cue for the step they represent.

References

1. C. Darwin, *On the Origin of Species*, 6th ed., John Murray, 1872.

2. F. Nietzsche, *Thus Spoke Zarathustra: A Book for All and None*, 1883–1885.

3. H. Spencer, *The Principles of Biology*, vol. 1, Williams & Norgate, 1864.

4. R. Dawkins, *The Selfish Gene*, Oxford University Press, 1976.