

SIMPLIPY Rule Discovery: the current algorithm (0.6.x)

The formal companion to docs/rules.md

Overview. Rules are mined, not hand-written: every valid source expression up to $L_{\max}$ tokens is enumerated (or sampled from the complete universe, per length), and each source is certified against a complete library of shorter candidates. A rule $\tau \rightarrow \tau'$ ships only if τ' reproduces τ 's values on a heavy-tailed, seeded evaluation matrix, survives an independent second verification, and is the *minimal* such target. The whole procedure is deterministic: one master seed reproduces the ruleset byte-for-byte, and a provenance sidecar records every parameter.

Algorithm 1 SIMPLIPY Rule Discovery

- 1: **Input:** operators $\mathcal{O}$, variables $\mathcal{V}$, literals $\mathcal{C}$; limits $L_{\max}$, L_{rep} ; master seed s : one `default_rng` derives X , X_{confirm} ($2 \times$ rows), and all mine/confirm seeds — byte-reproducible cross-machine; challenges $K=16$, retries $R=16$; tolerances $\text{rtol} = 10^{-9}$, $\text{atol} = 10^{-12}$
 - 2: $X \leftarrow$ heavy-tailed mixture: 40% $\mathcal{N}(0, 5)$, 25% $\mathcal{U}(-50, 50)$, 25% signed log-uniform $10^{-4}..10^3$, 10% exact specials $\{\pm 0, \pm 0.1, \pm 0.5, \pm 1, \pm 2, \pm e, \pm \pi, \pm 10\}$; evidence floor $m_{\min} = \text{rows}/8$
 - 3: Pre-compute $\mathcal{E}_L$ by bottom-up per-length dynamic programming (complete by construction); **assert** $|\mathcal{E}_L|$ equals the exact counting recurrence for every L , else **abort** the mine
 - 4: Universe policy per length: COMPLETE, an EXHAUSTIVE class slice, or SAMPLED (count-weighted uniform draws, validated; coverage always reported), with an optional saturation stop and a fixpoint re-mine of skipped draws' reducts
 - 5: **for** length L ascending (chunked with per-chunk checkpoints; per-source seed = length-seed + index, so chunking is bit-identical to one monolithic call) **do**
 - 6: $\mathcal{R}_{\text{new}} \leftarrow \emptyset$; candidate library built once, deterministically sorted, with the fold filter: variable-free candidates of length ≥ 2 dropped (the bare $\diamond$ candidate preempts them; behavior-preserving)
 - 7: **for** each $\tau \in \mathcal{E}_L$ (parallel, per-source seeds; the rule set is FIXED within a length and refreshed between lengths) **do**
 - 8: $\tau^* \leftarrow \text{SIMPLIPY}_{\mathcal{R}}(\tau)$
 - 9: **if** $|\tau^*| < |\tau|$ **then**
 - 10: **relaxed Kruskal:** still search, with the target bound tightened to $|\tau^*| - 1$ (strictly better targets only)
 - 11: **end if**
 - 12: All-constant sources: one shared exact leaf-value table (numeric literals *and* $\pi, e, -1$); variable-free sources collapse by exact interval value-class; interval reachability gate for variable-free candidates
 - 13: **for** length $j = 1$ to $\min(|\tau^*| - 1, L_{\text{rep}})$, candidates with $\text{Vars}(\tau') \subseteq \text{Vars}(\tau)$ **do**
 - 14: **if** $\text{Equivalent}^+(\tau, \tau')$ ▷ Algorithm 2 **then**
 - 15: track τ'_{best} by fewest $\diamond$; found $\leftarrow \text{True}$
 - 16: **end if**
 - 17: **end for**
 - 18: on found (wildcard-multiplicity check): add $\tau \rightarrow \tau'_{\text{best}}$ to $\mathcal{R}_{\text{new}}$
 - 19: **end for**
 - 20: **Stage-2 confirm:** re-verify every mined pair on the independent X_{confirm} with fresh constants and a per-rule content-derived seed (order-independent verdicts); rejects are dropped
 - 21: $\mathcal{R} \leftarrow \mathcal{R} \cup \text{DEDUPLICATE}(\mathcal{R}_{\text{new}})$ ▷ canonical wildcard patterns; shortest target per source
 - 22: Checkpoint rules + provenance sidecar (all parameters, derived seeds, X specification, per-length universe counts/coverage) and update the completeness marker — a mine is complete only when the marker equals the target length
 - 23: **end for**
 - 24: **return** $\mathcal{R}$
-

Algorithm 2 Expression Equivalence Check (Equivalent⁺)

```
1: Input:  $\tau, \tau', X$ ;  $K=16$  challenges,  $R=16$  retries;  $\text{rtol} = 10^{-9}$ ,  $\text{atol} = 10^{-12}$ ; evidence floor  $m_{\min}$ 
2: Source instances hoisted once per source: per-constant pole-grid sweep (0.02..500, including integers) +
   random magnitude rounds, sign combinations  $s \in \{-1, +1\}^{|C|}$ 
3: for each challenge instance do
4:   Generic equivalence with domain extension: rows where the source is finite must match within
      $\text{rtol}/\text{atol}$ ; rows where it is NaN/ $\pm\infty$  may be completed by the target; the reverse (domain restriction)
     rejects; a per-instance interval domain-preservation gate rejects rewrites inventing a function off the
     source's domain
5:   Evidence gate: require  $\geq m_{\min}$  unique source-finite rows (no certification without evidence)
6:   constant-free: f64 comparison is a pre-filter; near-misses are re-judged on exactly the failing rows at
     1024-bit precision
7:   constant-bearing: order-independent fit seed (source seed + content hash + instance); closed-form
     affine/log-linear fits (Householder QR) then Levenberg–Marquardt restarts; near-miss escalates through
     a robust median-trimmed refit, then high-precision re-judgment at the refit constants
8:   if  $\neg \text{match}$  then
9:     return False
10:  end if
11: end for
12: return True
```

Historical note (the 0.5.0 enumeration fix). The published algorithm assumes complete enumeration of $\mathcal{E}_L$. Versions $\leq 0.4.x$ instead grew the universe by whole-set composition passes that exited as soon as *one* maximal-length expression existed — silently enumerating only a fraction of the longer strata (with a 7-operator configuration: 6.4% of $\mathcal{E}_4$, 2.1% of $\mathcal{E}_5$, 0.085% of $\mathcal{E}_6$, 0.00089% of $\mathcal{E}_7$). This was a *coverage* defect, not a soundness defect: nothing wrong was certified, but most long sources were never examined. 0.5.0 replaced it with the bottom-up per-length DP above, asserted against the exact counting recurrence on every run.