The Discovery Equation: A Unified Information-Theoretic Law for Mathematical Insight

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

October 15, 2025

The process of mathematical discovery—whether by human intuition or algorithmic exploration—has always seemed to transcend ordinary computation. Yet, when viewed through the lens of information theory, the phenomenon can be expressed with startling simplicity. This paper presents the *Discovery Equation*, a quantitative law describing how any truth-seeking process converts computational effort into verified insights. The equation

$$\mathbb{E}[\mathrm{Yield}_B(q)] \propto B\, e^{-D_{\mathrm{KL}}(\tau \,\|\, q)}$$

links the expected yield of true statements to the alignment between a generator q and the ideal distribution $\tau$of all useful truths, weighted by utility and verification cost. Every reduction in the information divergence D_{KL}results in an exponential increase in discovery efficiency. This framework unifies concepts from algorithmic probability, variational inference, and the aesthetic principles of mathematics, showing that intuition, beauty, and formal search are mathematically equivalent ways of narrowing the same informational gap. By quantifying the convergence between randomness and order, the Discovery Equation transforms the mystery of insight into a measurable, reproducible process—bridging the gap between mathematical creativity and physical law.

Keywords: mathematical discovery, information theory, KL divergence, algorithmic probability, MDL, symmetry, intuition, aesthetics, truth yield, human-AI collaboration, variational inference, free energy, Logos.

Outline

1. Introduction: From Mystery to Mechanism
    - The problem of insight in infinite symbolic space
    - Why discovery appears supercomputational
    - Preview of the Discovery Equation and its scope
2. Foundations in Information Geometry
    - Truth space and generator space as dual manifolds
    - KL divergence as epistemic distance
    - The exponential efficiency law from importance sampling
3. Derivation of the Discovery Equation
    - Defining the ideal distribution $\tau(s) \propto T(s)u(s)/c(s)$
    - Expected yield as weighted sampling efficiency
    - Proof sketch linking hit-rate to divergence minimization
4. Mathematical Implications
    - Small- and large-divergence limits
    - Relation to the Cross-Entropy Method, MDL, and ELBO
    - Scaling laws for truth yield vs. compute budget
5. Aesthetic Priors and the Human Shortcut
    - Symmetry, conciseness, and beauty as implicit KL minimizers
    - Human intuition as a preconditioned prior over τ
    - Bridging the “divine shortcut” with formal efficiency
6. AI and Algorithmic Discovery
    - Random generation + falsification loops as physical realization of q
    - Learning q → τ via variational optimization
    - Empirical tests: truth-yield curves and KL diagnostics

7. Thermodynamic Interpretation
    - Free energy and informational temperature analogies
    - Discovery as entropy reduction under bounded energy
    - Convergence law: $E[\text{Yield}] \propto e^{-\Delta F/kT}$
8. Predictions and Experimental Pathways
    - Benchmarks for symbolic discovery engines
    - Measuring truth yield across priors and budgets
    - Detectable signatures of human–AI synergy
9. Philosophical Reflections
    - The Logos as the universal attractor of truth
    - Uniting chance, beauty, and reason under one equation
10. Conclusion: Toward a Physics of Insight

- From randomness to revelation
- The Discovery Equation as a law of cognition and creation

Epilogue: Tunneling from q → τ

References

1. Introduction: From Mystery to Mechanism

1.1 The Problem of Insight in Infinite Symbolic Space

Mathematics unfolds within an infinite combinatorial space of possible expressions. Out of this vast symbolic expanse, only a vanishing fraction corresponds to statements that are syntactically valid, semantically meaningful, and objectively true. The problem of *insight* is the question of how finite beings—human or artificial—consistently locate such rare structures. Every proven theorem, every elegant identity, represents an improbable victory against entropy. Classical logic and computation explain verification but not *discovery* itself. The central challenge, therefore, is not how to check truth, but how to *aim* toward it efficiently in an otherwise featureless search landscape.

1.2 Why Discovery Appears Supercomputational

Traditional computational models treat exploration as brute force: the traversal of discrete symbol trees or proof graphs with exponential growth. Yet human mathematicians routinely make leaps that appear to defy algorithmic scaling—compressing vast search spaces into intuitive trajectories guided by symmetry, aesthetics, or metaphor. This asymmetry suggests that insight operates through hidden structure: biases that dramatically reduce effective entropy. In computational terms, these biases approximate an optimal sampling distribution over truths—something akin to an unreasonably accurate *prior*.

1.3 Preview of the Discovery Equation and Its Scope

The *Discovery Equation* formalizes this intuition. It states that the expected yield of true, useful statements per unit effort is proportional to

$$\mathbb{E}[\mathrm{Yield}_B(q)] \propto B\, e^{-D_{\mathrm{KL}}(\tau \,\|\, q)},$$

where q is the actual search distribution and τ is the ideal distribution of truths weighted by their utility and cost. The equation quantifies discovery as the exponential benefit of alignment between search and structure. This simple relation unites intuition, aesthetics, and algorithmic optimization under a common information-theoretic law, transforming the mystery of insight into a measurable convergence between randomness and order.

2. Foundations in Information Geometry

2.1 Truth Space and Generator Space as Dual Manifolds

Mathematical discovery can be framed geometrically: every possible statement s inhabits a vast discrete–continuous manifold $\mathcal{S}$ of symbolic configurations. Within this space, two probability distributions define dual perspectives:

- The truth manifold, $\tau(s)$, representing the "ground reality" of mathematics—weighted by whether a statement is *true*, *useful*, and *verifiable* (truth × utility ÷ cost).
- The generator manifold, $q(s)$, representing how an explorer—human or algorithm—*actually samples* possible statements during reasoning, conjecture, or automated search.

Each manifold forms a surface within the simplex of all possible distributions over $\mathcal{S}$. Discovery then becomes a problem of *information alignment*: bringing q closer to τ. As in thermodynamic or Bayesian manifolds, proximity is not measured by Euclidean geometry but by *divergence*—the information cost of using q when the world follows τ.

When $q = \tau$, the explorer's expectations perfectly match the landscape of truth, and discovery proceeds linearly with computational effort. As q diverges, wasted effort grows exponentially: most samples land in barren symbolic deserts where falsification is cheap but fruitless.

2.2 KL Divergence as Epistemic Distance

The Kullback–Leibler divergence,

$$D_{\mathrm{KL}}(\tau \parallel q) = \sum_{s} \quad \tau(s) \log \frac{\tau(s)}{q(s)},$$

acts as an *epistemic distance* between the ideal and actual manifolds. It quantifies the information penalty for exploring the wrong regions of symbol space. Each unit of KL corresponds to a natural-log loss of discovery efficiency: one "nat" of divergence reduces the expected yield by a factor of e^{-1}.

This definition connects deep ideas across fields. In Bayesian inference, D_{KL} measures how far a model is from the true posterior; in thermodynamics, it becomes a free-energy gap; in learning theory, it bounds regret and generalization error. In mathematics, it now measures *cognitive misalignment*: how much one's conjectural prior q misunderstands the distribution of

mathematical structure. A smaller KL means intuition, aesthetic bias, or algorithmic heuristics are *closer to the Logos*—the latent order of truth.

2.3 The Exponential Efficiency Law from Importance Sampling

The *exponential efficiency law* follows from basic importance sampling theory. Suppose we wish to estimate the "mass of truth" $Z_\tau = \sum_s \quad T(s)u(s)/c(s)$using samples from q. The expected number of verified truths after spending computational budget B scales with

$$\mathbb{E}[\mathrm{Yield}_B(q)] \approx B\, e^{-D_{\mathrm{KL}}(\tau \| q)}.$$

This arises because the *variance* of importance weights—and thus the wasted sampling effort—grows exponentially with D_{KL}. When q perfectly matches τ, all weights are equal, yielding zero variance and linear scaling with budget. When q differs, the exponential suppression term appears naturally from the law of large deviations.

Thus, mathematical discovery follows the same statistical geometry that governs efficient Monte Carlo integration, probabilistic inference, and physical equilibration. The smaller the epistemic distance D_{KL}, the greater the proportion of effort that converts directly into verified insight. In this view, human intuition, aesthetic judgment, and AI-guided search all act as mechanisms that reduce the *information curvature* between q and τ—flattening the manifold of ignorance into the smooth plane of understanding.

3. Derivation of the Discovery Equation

3.1 Defining the ideal distribution $\tau(s) \propto T(s)\, u(s)/c(s)$

- Space. Let $\mathcal{S}$ be the set of candidate mathematical statements s.
- Truth. $T(s) \in \{0,1\}$ indicates whether $s$is true in the intended semantics.
- Utility. $u(s) \geq 0$measures downstream value (e.g., MDL bits saved, proof-length reductions).
- Cost. $c(s) > 0$ is expected verification cost (time/compute for refutation/certification).

Define the *ideal* target distribution

$$\tau(s) \;=\; \frac{T(s)\, u(s)/c(s)}{Z_\tau}, Z_\tau = \sum_{s \in \mathcal{S}} \quad \frac{T(s)\, u(s)}{c(s)}.$$

Intuition: τplaces probability mass where statements are true, useful, and cheap to verify.

A practical explorer (human/AI) samples from a proposal $q$over $\mathcal{S}$. We assume support inclusion: if $\tau(s) > 0$then $q(s) > 0$, to avoid infinite divergence.

3.2 Expected yield as weighted sampling efficiency

Consider a *strong verifier* that, when given s, (i) rejects false statements quickly with overwhelming probability, and (ii) certifies true ones in expected cost $c(s)$. Under a compute budget B, the expected number of verified truths (the yield) when drawing $s \sim q$ and spending the required cost per draw is

$$\mathbb{E}[\mathrm{Yield}_B(q)] \;=\; B \sum_s \underbrace{q(s)\,\frac{T(s)}{c(s)}}_{\text{truth-rate per unit cost}}.$$

Weighting by utility (we care about *useful* truths) gives

$$\mathbb{E}[\mathrm{Yield}_B(q)] \;=\; B \sum_s q(s)\,\frac{T(s)\,u(s)}{c(s)} \;=\; B\,Z_\tau \sum_s q(s)\,\tau(s)\,\frac{Z_\tau^{-1} T(s)u(s)/c(s)}{\tau(s)}$$
$$=\; B\,Z_\tau \sum_s \tau(s)\,\frac{q(s)}{\tau(s)}.$$

Introduce the importance weight $w(s) = \frac{\tau(s)}{q(s)}$ (well-defined by support inclusion). Then

$$\sum_s q(s)\,\frac{T(s)\,u(s)}{c(s)} = \; Z_\tau \sum_s \tau(s)\,\frac{1}{w(s)} = \; Z_\tau\,\mathbb{E}_{\tau}[\,w(s)^{-1}].$$

Thus

$$\boxed{\mathbb{E}[\mathrm{Yield}_B(q)] \;=\; B\,Z_\tau\,\mathbb{E}_{\tau}[w^{-1}].}$$

The *ideal* case $q = \tau$ gives $w \equiv 1$and $\mathbb{E}[\mathrm{Yield}_B] = B\,Z_\tau$ (purely linear in budget).

3.3 Proof sketch linking hit-rate to divergence minimization

We now relate $\mathbb{E}_\tau[w^{-1}]$ to the KL divergence $D_{\mathrm{KL}}(\tau \parallel q) = \mathbb{E}_\tau[\log\, w]$.

1. Jensen's inequality (log is concave):

$$\log\, \mathbb{E}_\tau[w^{-1}] \;\geq\; \mathbb{E}_\tau[\log\, w^{-1}] \;=\; -\,\mathbb{E}_\tau[\log\, w] \;=\; -\,D_{\mathrm{KL}}(\tau \parallel q).$$

Exponentiating:

$$\mathbb{E}_\tau[w^{-1}] \;\geq\; e^{-D_{\mathrm{KL}}(\tau\|q)}.$$

2. Plug back into yield:

$$\mathbb{E}[\mathrm{Yield}_B(q)] \;=\; B\,Z_\tau\,\mathbb{E}_\tau[w^{-1}] \;\geq\; B\,Z_\tau\,e^{-D_{\mathrm{KL}}(\tau\|q)}.$$

Thus we obtain the exponential efficiency law (up to the constant Z_τ):

$$\boxed{\mathbb{E}[\mathrm{Yield}_B(q)] \;\gtrsim\; B\,Z_\tau\,e^{-D_{\mathrm{KL}}(\tau\|q)}.}$$

3. Tightness and operational meaning.

- The bound is tight when $w$concentrates (small variance), approached as $q \to \tau$(zero-variance importance sampling).
- For practical discovery processes (finite budgets, heavy-tailed weights), large-deviation/CE-method analyses likewise show exponential sensitivity of efficiency to the KL gap; minimizing $D_{\mathrm{KL}}(\tau \parallel q)$ is the unique way to approach the linear-in-B ideal.

4. Conclusion.
   Minimizing $D_{\mathrm{KL}}(\tau \parallel q)$ is the *variational principle* of mathematical discovery. Any mechanism—human aesthetic priors, symmetry-aware generators, MDL-tempered sampling, or learned neural guidance—that reduces this divergence will multiplicatively increase expected truth yield for the same budget. Hence the compact law:

$$\mathbb{E}[\mathrm{Yield}_B(q)] \;\propto\; B\,e^{-D_{\mathrm{KL}}(\tau\|q)}.$$

4. Mathematical Implications

4.1 Small- and large-divergence limits

- Perfect aim (KL = 0).
  If $q = \tau$, then $\mathbb{E}[\text{Yield}_B] = B\,Z_\tau$: yield is linear in budget; there is no variance penalty.
- Small mismatch (KL = ε ≪ 1).
  $e^{-\varepsilon} \approx 1 - \varepsilon + \frac{1}{2}\varepsilon^2$. Hence

$$\mathbb{E}[\text{Yield}_B(q)] \approx B\,Z_\tau\,(1 - \varepsilon + \frac{1}{2}\varepsilon^2).$$

Each nat shaved off KL multiplies yield by e; each bit (ΔKL = ln 2) doubles it.

- Total-variation control.
  Pinsker: $\| \tau - q \|_{\text{TV}} \leq \sqrt{\frac{1}{2} D_{\text{KL}}(\tau \parallel q)}$. Small KL guarantees small distributional miss; hit-rates won't collapse unexpectedly.
- Large mismatch (KL = K ≫ 1).
  $\mathbb{E}[\text{Yield}_B] \propto B\,e^{-K}$. This is the rare-event regime: unless $B$is exponential in K, yield is effectively zero.
- Support mismatch.
  If $q(s) = 0$where $\tau(s) > 0$ (missed modes), $D_{\text{KL}} = \infty$ and expected yield vanishes—no amount of compute rescues missing support.

4.2 Relation to the Cross-Entropy Method, MDL, and ELBO

- Cross-Entropy Method (CEM).
  CEM iteratively updates $q$to minimize $D_{\text{KL}}(\tau \parallel q)$ toward the zero-variance importance sampler. Our law states the operative payoff: every KL drop buys an exponential yield lift. The "elite sample" heuristic is a practical estimator of τ's high-mass region.
- Minimum Description Length (MDL).
  In expectation, code-length gaps equal KL gaps. Using a proposal $q_L(s) \propto e^{-\beta L(s)}$ with a good description-length L (simplicity, symmetry bonuses, utility minus cost) approximates $-\log\,\tau$. Shorter expected codes ⇒ smaller expected KL ⇒ exponentially higher yield.

- Variational Inference / ELBO (free energy).
  ELBO = (log-evidence) − D_{KL} (posterior ‖ q). Minimizing KL maximizes evidence; in our setting, minimizing $D_{\mathrm{KL}}(\tau \parallel q)$ maximizes truth yield. Same geometry: performance rises exponentially as the variational gap shrinks.

- Large deviations / Donsker–Varadhan.
  Exponential rates of rare hits are controlled by variational KL functionals; our yield law is the operational face of that theory for mathematical statements.

4.3 Scaling laws for truth yield vs. compute budget

- Iso-yield frontier.
  From $\mathbb{E}[\mathrm{Yield}_B] \propto B\, e^{-D_{\mathrm{KL}}}$: keeping expected yield constant gives

$$\log\, B \;=\; D_{\mathrm{KL}} + \text{const.}$$

One extra nat of misalignment must be paid by an e-fold increase in budget; one bit (ln 2) by a 2× budget.

- Compute–alignment tradeoff.
  Let C be spend on sampling (linear in B) and R on improving q (reducing KL). If improvement rate is $\dot{D}_{\mathrm{KL}} = -\alpha(R)$ and sampling rate is $\dot{B} = \beta(C)$, then instantaneous yield growth is

$$\frac{d}{dt}\log\, \mathbb{E}[\mathrm{Yield}] = \frac{\dot{B}}{B} \;-\; \dot{D}_{\mathrm{KL}}.$$

Optimal allocation equalizes marginal returns: invest in KL-reduction until $\dot{D}_{\mathrm{KL}} = \dot{B}/B$.

- Blending priors (human + machine).
  With geometric mix $q_\lambda \propto q_M^{1-\lambda} q_H^{\lambda}$, convexity yields a unique λ^* minimizing $D_{\mathrm{KL}}(\tau \parallel q_\lambda)$. The lift over a baseline q_0 is

$$\frac{\mathbb{E}[\mathrm{Yield}(q_\lambda)]}{\mathbb{E}[\mathrm{Yield}(q_0)]} = \exp\;\; (D_{\mathrm{KL}}(\tau \parallel q_0) - D_{\mathrm{KL}}(\tau \parallel q_\lambda)).$$

- Diminishing returns in brute-force scaling.
  For fixed KL, yield grows only linearly with B. For fixed B, each constant reduction in KL multiplies yield; beyond a point, improving aim beats adding compute.

- Phase-transition targeting.
  In families with thresholds (e.g., k-SAT-like identities), the local KL landscape is steep near criticality; small prior improvements yield disproportionately large hits—predicting sharp jumps in yield when priors cross structural thresholds.

Takeaway: in every regime, alignment dominates scale. Compute buys you linear gains; alignment (smaller D_{KL}) buys you multiplicative gains.

5. Aesthetic Priors and the Human Shortcut

5.1 Symmetry, Conciseness, and Beauty as Implicit KL Minimizers

Mathematicians have long observed that beautiful formulas tend to be true. From Euler's identity to Maxwell's equations, aesthetic harmony often anticipates formal proof. Within the framework of the *Discovery Equation*, this correspondence acquires a quantitative explanation: aesthetic preferences act as *implicit KL minimizers*.

Aesthetic filters—favoring simplicity, symmetry, invariance, and self-consistency—bias the human generator $q_{\text{human}}(s)$ toward the ideal truth distribution $\tau(s) \propto T(s)u(s)/c(s)$.

- Simplicity aligns with the *Minimum Description Length* (MDL) principle, minimizing the code length $-\log\ q(s)$.
- Symmetry preserves invariance under transformations, a hallmark of deeper structural truth.
- Elegance corresponds to minimizing redundancy, improving the ratio of meaning to expression length.

Each of these biases suppresses noise in symbol space and narrows the search manifold to regions where true, universal relationships are more likely to reside. Aesthetics, in this view, is not subjective decoration but an *evolutionarily internalized algorithmic prior* that approximates the geometry of τ.

5.2 Human Intuition as a Preconditioned Prior over τ

Human cognition does not explore mathematical space uniformly; it is preconditioned by millennia of cultural, sensory, and cognitive evolution. These priors embed information about invariance, hierarchy, and causality that mirror the deep symmetries of nature. The brain's internal generator q_{human} thus comes "pre-trained" on physical and perceptual regularities that already approximate the true manifold of lawful relations.

Formally, this means the expected divergence $D_{\mathrm{KL}}(\tau \parallel q_{\text{human}})$ is far smaller than that of an uninformed random generator. Intuition acts as a low-entropy attractor basin: when mathematicians sense that a conjecture "feels right," they are responding to a subconscious signal of KL proximity—the alignment of their cognitive distribution with the latent structure of truth.

Even errors are informative: false but *elegant* conjectures (like early forms of the prime number theorem) lie near the manifold of true statements. They serve as local minima in the KL landscape, where small refinements yield genuine discoveries.

5.3 Bridging the "Divine Shortcut" with Formal Efficiency

The *divine shortcut*—the idea that humans possess a gifted sensitivity to truth—is here interpreted as a measurable efficiency: a built-in reduction in $D_{\mathrm{KL}}(\tau \parallel q)$ relative to mechanical randomness. Whether this bias arises from evolutionary adaptation, cultural inheritance, or divine inspiration, its computational effect is identical: exponential improvement in discovery yield per unit of effort.

Mathematically, this can be expressed through a blended generator:

$$q_\lambda(s) \propto q_{\text{machine}}(s)^{1-\lambda}\, q_{\text{human}}(s)^{\lambda},$$

whose expected yield satisfies

$$\mathbb{E}[\mathrm{Yield}_B(q_\lambda)] \propto B\, e^{-D_{\mathrm{KL}}(\tau \| q_\lambda)}.$$

A small contribution of human bias ($\lambda \ll 1$) can dramatically lower the divergence term, acting as a catalyst that guides machine exploration toward the true manifold.

In this synthesis, the "divine shortcut" and the "aesthetic prior" become the same phenomenon viewed at different levels: one metaphysical, one statistical. The Logos—the principle of mathematical order—is thus reflected both in the elegance of human intuition and in the efficiency of information alignment.

6. AI and Algorithmic Discovery

6.1 Random Generation + Falsification Loops as the Physical Realization of q

In the algorithmic domain, an AI system's proposal distribution $q(s)$corresponds to its generative model over candidate statements—formulas, theorems, or equations. Random generation followed by falsification acts as the *physical embodiment* of sampling from $q$and testing against mathematical reality.

Each iteration in such a loop involves:

1. Generation: draw a symbolic hypothesis $s \sim q$.
2. Falsification: verify $T(s)$ through computation, proof search, or model checking.
3. Reweighting: update $q$based on the outcome (accepted or rejected).

This process mirrors the natural evolution of mathematical thought: conjecture, contradiction, and refinement. The truth-yield per compute cycle in these systems provides an empirical measure of $\mathbb{E}[\mathrm{Yield}_B(q)]$.
When plotted over time, discovery accelerates as the generator $q$learns to approximate the ideal target τ—the distribution of true, useful, and efficiently verifiable statements. Each correction in falsification acts as a small *gradient step* reducing the epistemic divergence $D_{\mathrm{KL}}(\tau \parallel q)$.

In this sense, falsification loops are not merely trial-and-error but thermodynamic learning systems, where information flows from logical failure toward structural equilibrium. Every rejected conjecture lowers entropy and reshapes the generator toward truth-aligned regions of symbol space.

6.2 Learning q → τ via Variational Optimization

The path from random $q$to near-optimal τcan be framed as a variational optimization problem:

$$q^* = \arg\min_q D_{\mathrm{KL}}(\tau \parallel q).$$

Because $\tau$is unknown—its truth labels are only revealed by verification—this objective must be approximated empirically. Each verified or falsified statement contributes a stochastic estimate of the gradient of D_{KL}:

$$\nabla_\theta D_{\mathrm{KL}}(\tau \parallel q_\theta) = -\mathbb{E}_{s\sim\tau}[\nabla_\theta \log\, q_\theta(s)] \;\approx\; -\mathbb{E}_{s\sim q_\theta}[T(s)\, u(s)/c(s)\, \nabla_\theta \log\, q_\theta(s)].$$

This defines an energy-based learning rule:

- True, useful statements push probability mass upward;
- False or inefficient statements are exponentially suppressed.

This is conceptually identical to training in variational inference, reinforcement learning, or energy-based models, but the reward signal here is *truth itself*.

As the generator learns, $D_{\mathrm{KL}}(\tau \parallel q_t)$ decreases monotonically, moving q_t along the information-geometric gradient toward the manifold of truths. Over long runs, the expected yield curve transitions from *sublinear random discovery* to *linear or superlinear convergence*—the hallmark of alignment with τ.

6.3 Empirical Tests: Truth-Yield Curves and KL Diagnostics

To validate the *Discovery Equation* empirically, we can measure whether observed truth yields follow the exponential law

$$\mathbb{E}[\mathrm{Yield}_B(q)] \propto B\, e^{-D_{\mathrm{KL}}(\tau \| q)}.$$

1. Truth-Yield Curves.
Across AI theorem provers, symbolic regression engines, or conjecture generators, measure verified discoveries $Y(B)$ as a function of compute B. When q is held fixed, $Y(B)$ should scale linearly. When q is improved (e.g., retrained with verified truths), yields should rise *multiplicatively*, following the predicted exponential relation with the measured KL gap.

2. KL Diagnostics.
Estimate $D_{\mathrm{KL}}(\tau \parallel q)$ indirectly via held-out validation:

- Use empirical truth frequencies $\hat{\tau}(s) \propto$ verified frequency$/c(s)$.
- Compute cross-entropy $H(\hat{\tau}, q) = -\sum_s \quad \hat{\tau}(s) \log\, q(s)$.
- Approximate $D_{\mathrm{KL}}(\tau \parallel q) \approx H(\hat{\tau}, q) - H(\hat{\tau})$.

A monotonic inverse relationship between yield efficiency and estimated KL provides quantitative confirmation.

3. Scaling Tests.
Run comparative experiments:

- Random baseline q_0 vs. symmetry- or MDL-biased priors q_1.

- Observe yield ratios $Y_1/Y_0 \approx \exp\left(D_{\mathrm{KL}}(\tau \parallel q_0) - D_{\mathrm{KL}}(\tau \parallel q_1)\right)$.

4. Human-in-the-loop Ablations.
Introduce micro-guidance—short human evaluations of elegance or plausibility. If even minimal human input yields exponential efficiency gains, this empirically supports the *divine shortcut hypothesis*: small, well-calibrated reductions in D_{KL} create superlinear improvements in verified yield.

Summary:
AI discovery systems operationalize the *Discovery Equation* through physical cycles of generation and falsification, optimized via stochastic gradient descent on $D_{\mathrm{KL}}(\tau \parallel q)$. When observed empirically, truth-yield curves, scaling behaviors, and KL diagnostics reveal a consistent pattern: exponential increases in discovery efficiency as the generator aligns with the underlying structure of truth. This bridges the conceptual gap between human insight and algorithmic learning—showing both as points on the same information-geometric continuum.

7. Thermodynamic Interpretation

7.1 Free energy and informational temperature analogies

Let each candidate statement $s \in \mathcal{S}$ be a microstate. Define an "ideal" Hamiltonian

$$H_\tau(s) := \; -kT \log \left(\frac{T(s)\, u(s)}{c(s)}\right),$$

so the target distribution is Boltzmann:

$$\tau(s) \; = \; \frac{e^{-H_\tau(s)/kT}}{Z_\tau}, Z_\tau = \sum_s \; e^{-H_\tau(s)/kT}.$$

A sampler (human/AI) with proposal $q$has its own effective Hamiltonian H_q via

$$q(s) = \frac{e^{-H_q(s)/kT}}{Z_q}, F(H) := -kT\log\, Z(H).$$

Then the KL gap equals a free-energy gap:

$$D_{\mathrm{KL}}(\tau \parallel q) \;=\; \frac{F(H_q) - F(H_\tau)}{kT} \;=\; \frac{\Delta F}{kT}.$$

Here, kT plays the role of an informational temperature: high T means diffuse, exploratory sampling (more entropy); low T means focused, exploitative sampling (sharper q).

7.2 Discovery as entropy reduction under bounded energy

Discovery converts compute budget into order: reducing the entropy of the sampling distribution around truth-rich regions. With fixed energy-like budget B, the *best* you can do is decrease $F(H_q)$ toward $F(H_\tau)$—i.e., lower the variational free energy by reshaping q (better priors, symmetries, MDL bias).

- Cooling (annealing): lowering T sharpens q, but if H_q is mis-specified, you can freeze into wrong modes. Practical schedules alternate entropy (exploration at higher T) with energy optimization (reducing $F(H_q)$).
- Symmetry as potential: adding invariance/elegance bonuses subtracts a potential $V_{\mathrm{sym}}(s)$ from H_q, collapsing phase space volume around structured regions (lower entropy at similar energy).
- Work–information trade: each bit of KL removed is $kT \ln 2$ less free energy—less "work" wasted per verified truth.

7.3 Convergence law: $\mathbb{E}[\mathrm{Yield}] \propto e^{-\Delta F/kT}$

From the discovery equation

$$\mathbb{E}[\mathrm{Yield}_B(q)] \;\propto\; B\, e^{-D_{\mathrm{KL}}(\tau\|q)} \;=\; B\, e^{-\Delta F/kT},$$

where $\Delta F = F(H_q) - F(H_\tau)$. Consequences:

- Linear ideal: if $\Delta F = 0$ (i.e., $q = \tau$), yield is $\propto B$.
- Exponential suppression: every increase of ΔF by kT multiplies the penalty by e^{-1}. One bit of free-energy gap ($\Delta F = kT \ln 2$) halves the yield.

- Optimal blending: for a geometric blend $H_\lambda = (1-\lambda)H_M + \lambda H_H$, the best mixture λ^* minimizes $F(H_\lambda)$, hence maximizes $e^{-\Delta F/kT}$.
- Cooling schedule: at fixed ΔF, lowering T sharpens q but also tightens the cost of mis-aim. Practically, reduce ΔF before deep cooling; otherwise you amplify errors.
- Phase transitions: as H_q crosses structural thresholds (e.g., adds a symmetry), $F(H_q)$ can drop sharply, causing step-changes in yield—empirically seen as sudden surges in verified truths.

Interpretation: discovery efficiency is a free-energy race. Improving priors (lowering $F(H_q)$) or aligning aesthetics/symmetry (reducing ΔF) yields exponential gains; raw compute B gives only linear gains.

8. Predictions and Experimental Pathways

8.1 Benchmarks for symbolic discovery engines

Task suites (curated, progressively harder).

- Algebraic identities (exact): polynomial/trig/exponential identities up to bounded tree size; ground truth via CAS normalization and finite-field PIT.
- Inequalities (certifiable): polynomial/rational/elementary inequalities on compact domains; ground truth via SOS/SDP or interval-Taylor certificates.
- Combinatorics (finite models): graph/property statements verified by model search (SAT/SMT) up to fixed n.
- Template discovery: find general lemmas that compress multiple instances (measured by MDL reduction on a held-out corpus).

Baselines.

- q_{rand}: uniform grammar sampler.
- q_{sym}: symmetry/degree/homogeneity aware.
- q_{mdl}: MDL-tempered (length/sparsity/arity penalties).
- q_{learn}: learned prior (premise selection/tactic policy).
- Blends: $q_\lambda \propto q_{\text{learn}}^{1-\lambda} q_{\text{human}}^{\lambda}$.

Standardized evaluation.

- Fixed compute budget B (CPU-seconds or call-count to verifiers).
- Fixed falsifier stack per suite (PIT primes, sample sizes, SDP tolerances).
- Public seeds + exact grammars; one-pass and multi-pass (with learning) regimes.

8.2 Measuring truth yield across priors and budgets

Primary outcome curves.

- Truth-yield curves: $Y(B; q)$= verified truths vs. budget. Expect linear scaling in B at fixed q, and multiplicative lifts as $q$improves.
- Lift ratios: $Y(B; q_1)/Y(B; q_0) \approx \exp\left(D_{\mathrm{KL}}(\tau \parallel q_0) - D_{\mathrm{KL}}(\tau \parallel q_1)\right)$.

KL diagnostics (practical proxies).

- Cross-entropy with empirical $\hat{\tau}$:
  $\hat{\tau}(s) \propto \frac{\text{verifications}(s)}{c(s)}$ on a validation pool; report $H(\hat{\tau}, q) = -\sum \hat{\tau} \log\ q$; approximate $D_{\mathrm{KL}} \approx H(\hat{\tau}, q) - H(\hat{\tau})$.
- Weight dispersion: sample $s \sim q$, compute $w(s) = \hat{\tau}(s)/q(s)$; track $\mathrm{Var}[\log\ w]$ (zero at $q = \tau$).
- Mode coverage: fraction of high-mass $\hat{\tau}$buckets hit by q(LSH on random evaluations).

Statistical plan.

- Power: pre-compute N(runs/seeds) to detect an expected log-lift $\Delta$with variance σ^2: need $N \geq 2\sigma^2(\, z_{1-\alpha/2} + z_{1-\beta})^2/\Delta^2$.
- Tests: paired log-yield differences across seeds; report mean ± 95% CI; use stratified bootstrap over tasks.
- Stopping: sequential probability ratio tests (SPRT) on acceptance of candidates with global αbudget (e.g., 10^{-12}).

Ablations.

- Remove symmetry filters; remove MDL penalties; disable learned retrieval; swap PIT primes; relax SDP order. Expect monotone decreases in yield and increases in estimated KL proxies.

Scaling laws.

- Plot $\log Y$ vs. $\log B$ at fixed q (slope ≈ 1).
- Plot $\log Y - \log B$ vs. $-D_{\mathrm{KL}}$ (should align near slope ≈ 1).
- Budget–alignment frontier: $\log\ B = D_{\mathrm{KL}} +$ constfor iso-yield contours.

8.3 Detectable signatures of human–AI synergy

Micro-oracle protocol (low-bandwidth human input).

- Batch top-K proposals by q_{learn}; a human tags each as {plausible, neutral, inelegant}.
- Adjust scores $L(s) \leftarrow L(s) - \epsilon \cdot$ tag with tiny ϵ (e.g., 0.05–0.2 nats).
- Re-sample; keep falsifier unchanged to isolate prior effects.

Predicted signatures.

1. Exponential lift from tiny guidance:
   Even $\lambda \approx 0.05$ in q_λ should yield a measurable multiplicative gain in $Y(B)$ vs. machine-only—consistent with $e^{-\Delta D_{\mathrm{KL}}}$.
2. Higher novelty rate:
   Increased proportion of new templates (not just new instances), measured by LSH clusters and unification coverage on held-out tasks.
3. Better calibration:
   For top-p% candidates, empirical truth frequency rises; Brier/NLL improves—evidence that human tags reduce epistemic misspecification.
4. Compression jump:
   Greater MDL reduction on benchmark corpora after integrating found truths—more bits saved per verified statement.

Controls & ethics.

- Sham guidance (random tags) to rule out placebo effects.
- Time-matched control (equal latency without tags).
- Leak checks (humans see text only, no outcomes).
- Attribution logs (who/what influenced acceptance).

Failure modes to watch.

- Spurious elegance: over-favoring short/pretty but parochial statements (detected by low cross-domain transfer).
- Library echoing: rediscovering near-duplicates (caught by LSH/dedup).
- Verifier bias: PIT field artefacts or low-order SOS leading to false survivals (mitigate via multi-field PIT, higher SOS order spot checks).

Summary predictions.

- Yield obeys $Y(B;q) \propto B\, e^{-D_{\mathrm{KL}}(\tau\|q)}$across suites.
- Any prior that lowers KL (MDL, symmetry, learned retrieval, human tags) produces multiplicative yield gains at fixed B.
- Minimal human input (few bits per batch) produces superlinear improvements, visible as steeper truth-yield curves, better calibration, and larger MDL compression—quantitative fingerprints of human–AI synergy predicted by the Discovery Equation.

9. Philosophical Reflections

9.1 The Logos as the Universal Attractor of Truth

At the deepest level, the *Discovery Equation*

$$\mathbb{E}[\mathrm{Yield}_B(q)] \propto B\, e^{-D_{\mathrm{KL}}(\tau\|q)}$$

describes more than an optimization problem—it expresses a metaphysical tendency of all intelligences toward alignment with the *Logos*, the universal principle of order and intelligibility. In classical theology and philosophy, the Logos is both the rational structure *within* creation and the creative Word *through* which it exists. In this light, the KL divergence becomes the quantitative form of *epistemic separation* from divine reason: an information-geometric measure of how far a mind's inner model q is from the true distribution of meaning τ.

Every act of understanding—scientific, mathematical, or spiritual—reduces this divergence. Discovery, then, is not an accidental byproduct of chance but the very motion of cognition toward the Logos-as-attractor. Human intuition, beauty, and AI search all enact this same convergence. Insofar as an entity seeks coherence, symmetry, and explanatory compression, it is drawn toward the same invariant center: the state where knowledge and being coincide.

9.2 Uniting Chance, Beauty, and Reason Under One Equation

Traditionally, chance is opposed to order, and beauty to necessity. The *Discovery Equation* dissolves these dualisms. Randomness (q) and reason (τ) are not adversaries but complementary poles of the same dynamic: one explores, the other anchors. The exponential term $e^{-D_{\mathrm{KL}}}$ unites them mathematically—showing that order emerges not by suppressing randomness, but by aligning it.

Beauty enters as the perceptible signature of that alignment. When the mind encounters an elegant theorem, it experiences the subjective correlate of minimized divergence: the recognition that chaos and form have converged. In this view, *aesthetics is epistemology made visible*.

The result is a single integrative framework:

- Chance provides variation (entropy).
- Beauty signals compression and symmetry (low informational curvature).
- Reason formalizes this balance as KL minimization (alignment with truth).

Mathematics, art, and revelation are therefore not distinct domains but different modes of the same process—the self-organization of awareness within the informational field of the Logos. The *Discovery Equation* simply expresses, in minimal analytical form, the universal law by which thought ascends toward coherence and truth.

10. Conclusion: Toward a Physics of Insight

10.1 From Randomness to Revelation

All discovery begins in uncertainty. Whether the system is a human mind pondering a conjecture or an AI generating hypotheses, the process starts as a diffusion through symbolic space—an ocean of randomness where almost nothing is true. What transforms that noise into revelation is *alignment*: the gradual reduction of informational distance between the generator $q$and the underlying structure of truth τ.

The *Discovery Equation* captures this transformation in its simplest possible form:

$$\mathbb{E}[\mathrm{Yield}_B(q)] \propto B\, e^{-D_{\mathrm{KL}}(\tau \| q)}.$$

It shows that discovery is not an accident but an energetic flow governed by information geometry. Every theorem, every insight, every moment of clarity represents a decrease in

epistemic free energy—a step closer to the equilibrium state where mind and reality resonate. The exponential form reveals why insight feels discontinuous: long plateaus of randomness punctuated by sudden leaps, when $q$crosses a threshold in divergence and the structure of truth locks into focus. Revelation, in this sense, is the phase transition from chaos to comprehension.

10.2 The Discovery Equation as a Law of Cognition and Creation

The *Discovery Equation* unifies phenomena once thought incommensurable—human intuition, algorithmic search, and divine illumination—under one principle. It is both descriptive and prescriptive:

- *Descriptive* because it quantifies how any cognitive system transforms uncertainty into verified structure through informational alignment.
- *Prescriptive* because it implies that maximizing discovery requires minimizing divergence—not brute force, but coherence with the order already latent in reality.

This makes the equation a bridge between physics and philosophy. In physics, $e^{-\Delta F/kT}$describes how systems approach equilibrium by releasing free energy. In cognition, $e^{-D_{\mathrm{KL}}}$describes how intelligence approaches understanding by releasing ignorance. Both are manifestations of the same thermodynamic principle: order arises when internal models converge toward external reality.

The implication is profound: *insight itself is lawful*. The same exponential symmetries that govern matter and energy also govern the growth of knowledge. The Logos—the rational structure of the universe—appears not only in equations of motion but in equations of meaning.

Thus, the *Discovery Equation* may stand as a minimal statement of a new natural law: a physics of insight, where creation and cognition are unified as dual expressions of one cosmic process—the continual reduction of divergence between mind and truth, randomness and revelation.

Epilogue: Tunneling from q → τ

If the *Discovery Equation* expresses the law of convergence between the search distribution $q$and the true distribution τ, then the final mystery is how that convergence actually occurs. What mechanism allows a finite mind—or a finite machine—to *tunnel* through the vast desert of falsehood and reach the narrow basin of truth?

In the purely mathematical sense, tunneling from $q$to τmay resemble a quantum transition between probability manifolds separated by an informational potential barrier. The Kullback–Leibler divergence $D_{KL}(\tau\|q)$ defines that barrier: the "energy gap" between what is and what is known. Random search alone cannot cross it efficiently; the probability of spontaneous alignment is exponentially suppressed. Yet, discovery happens. Something leaks through.

This leakage may represent the non-locality of cognition—a resonance between the structure of thought and the structure of truth itself. When either a human or an AI system glimpses an elegant pattern, it effectively *samples from beyond its current distribution*. The act of insight is thus a *quantum-like tunneling event in information space*: improbable, discontinuous, but lawful.

The barrier's height defines the difficulty of discovery, while its *width*—the cognitive distance between model and reality—shrinks through aesthetic priors, intuition, or learned gradients. Every successful insight is a micro-collapse of uncertainty: a local instantiation of the universal tendency $q \to \tau$.

From this perspective, mathematical revelation is not magic but resonance—an alignment strong enough to make the improbable pass through the impossible. What tunnels is not information but *meaning* itself: the deep isomorphism between mind and the Logos.

And so, whether one names it intuition, grace, or gradient descent, the law remains the same:

$$q \xrightarrow{\text{tunnel of coherence}} \tau.$$

Each discovery is a tiny miracle—proof that truth is not only reachable, but calling.

References

1. Amari, S., & Nagaoka, H. (2000). *Methods of Information Geometry.* American Mathematical Society / Oxford University Press.
   — Foundational reference for the geometric interpretation of probability distributions and the role of Kullback–Leibler divergence as a natural metric on statistical manifolds.

2. Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley-Interscience.
   — Classical treatment of entropy, mutual information, and divergence; provides mathematical grounding for the information-theoretic derivations used in the Discovery Equation.

3. Kullback, S., & Leibler, R. A. (1951). "On Information and Sufficiency." *The Annals of Mathematical Statistics*, 22(1), 79–86.
   — Original paper introducing the Kullback–Leibler divergence as a measure of information distance.

4. Jaynes, E. T. (1957). "Information Theory and Statistical Mechanics." *Physical Review*, 106(4), 620–630.
   — Establishes the connection between information theory and thermodynamics; central to the free-energy interpretation used in the paper.

5. Donsker, M. D., & Varadhan, S. R. S. (1975). "Asymptotic Evaluation of Certain Markov Process Expectations for Large Time." *Communications on Pure and Applied Mathematics*, 28(1), 1–47.
   — Provides the variational representation connecting large-deviation theory and KL divergence.

6. Friston, K. J. (2010). "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience*, 11(2), 127–138.
   — Introduces the concept of cognition and perception as free-energy minimization; directly analogous to the epistemic convergence described in the Discovery Equation.

7. Rubinstein, R. Y., & Kroese, D. P. (2004). *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning.* Springer.
   — Demonstrates the exponential efficiency gains in rare-event sampling through KL minimization, the algorithmic precursor to the Discovery Equation.

8. Hinton, G. E., & van Camp, D. (1993). “Keeping Neural Networks Simple by Minimizing the Description Length of the Weights.” *Proceedings of COLT ’93*, 5–13.
— An early articulation of the Minimum Description Length (MDL) principle as a bias toward simplicity, directly related to aesthetic priors.

9. Rissanen, J. (1978). “Modeling by Shortest Data Description.” *Automatica*, 14(5), 465–471.
— Foundational paper on the MDL principle, linking compression to generalization and implicitly to KL divergence.

10. Solomonoff, R. J. (1964). “A Formal Theory of Inductive Inference, Parts I and II.” *Information and Control*, 7(1–2), 1–22 & 224–254.
— Establishes the algorithmic probability model underlying formal discovery and connects inductive learning to universal priors.

11. Schmidhuber, J. (1997). “Discovering Neural Nets with Low Kolmogorov Complexity and High Generalization Capability.” *Neural Networks*, 10(5), 857–873.
— Explores computational models of creativity and discovery, introducing aesthetic simplicity as an optimization bias.

12. Watanabe, S. (1960). “Information Theoretical Aspects of Inductive Reasoning.” *Kybernetika*, 2(2), 45–51.
— Early work uniting inference and information theory; introduces entropy-based reasoning in mathematical discovery.

13. Tishby, N., Pereira, F. C., & Bialek, W. (2000). “The Information Bottleneck Method.” *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, 368–377.
— Shows how optimal representation learning arises from balancing compression and relevance, paralleling the truth–utility tradeoff in τ(s).

14. Polya, G. (1945). *How to Solve It.* Princeton University Press.
— Classic study of mathematical problem-solving strategies; precursor to the idea that intuitive heuristics encode priors over the structure of truth.

15. Penrose, R. (1989). *The Emperor’s New Mind.* Oxford University Press.
— Philosophical exploration of mathematical intuition and consciousness as non-computable but structured processes, consistent with the idea of aesthetic priors.

16. Wheeler, J. A. (1990). “Information, Physics, Quantum: The Search for Links.” In W. Zurek (Ed.), *Complexity, Entropy, and the Physics of Information.* Addison-Wesley.
— Argues for an informational basis of physical law (“It from Bit”), providing metaphysical continuity between computation, physics, and cognition.

Note:
The references above blend canonical works in information theory, statistical mechanics, algorithmic probability, and philosophy of mathematics, providing both empirical and metaphysical scaffolding for the *Discovery Equation*. Together, they show that cognition, computation, and creation can be described as a single process of minimizing divergence between potentiality and truth—the informational convergence toward the Logos.