

Toward a Predictive Theory of Scientific Progress

Which predictions can the literature support, and which remain conjectures?

Complexity, Information, and Cross-Domain Transfer

Why investigate a bottleneck instead of only a better generator?

Suppose more proposals bring no increase in independently validated discoveries. Poor candidates are one explanation; too few validation slots or repeated proposals are others.

Following a candidate through the process changes the question: where does additional effort stop producing useful evidence? That diagnosis suggests competing interventions.

Construct competing interventions from the same symptom

Candidate intervention	Evidence that would favor it
Improve the generator	Invalid or poor candidates dominate tested failures
Increase validation capacity	Promising unique candidates remain untested
Improve ranking / remove duplicates	Repeated or misranked candidates consume slots

No branch is selected by proposal count alone. Record unique candidates, independent validity, resource use and uncertainty.

Why cross two budgets rather than compare two runs?

A before/after comparison can change the model, tasks and evaluator together. Increasing proposals alone cannot reveal whether their benefit depends on validation capacity.

Cross low/high proposal budgets with low/high validation capacity on the same tasks. This creates the contrast that the bottleneck explanation predicts.

A feasible study separates search from validation

Same task, model and endpoint	20 validations	40 validations
Low proposal budget	Baseline yield	Capacity effect at low budget
High proposal budget	Search effect at low capacity	Combined effect / interaction

What result would change the explanation?

If larger pools help only with 40 validations, downstream capacity limits the observed search gain. If they help at both capacities, candidate quality or diversity may have improved.

If proxy scores rise but independent validation does not, proxy improvement has failed to establish discovery improvement. A null result needs an uncertainty interval.

A conditional target for prediction

$$Q(\pi, b; D, K) = \Pr(\text{validated outcome by } b)$$

For a specified task population D and knowledge state K , predict the probability of a validated outcome by resource vector b .

Keep the acceptance rule fixed so that a resource comparison measures progress toward the same endpoint.

A worked model: proposals, screening and assays

$$n \leq B/(1+10s); \quad E[Y]=nsv$$

Main endpoint: acceptance by one prespecified assay. Each proposal costs 1 unit; each tested screen-passer costs 10. Screen pass probability is s ; assay acceptance given a pass is v .

Expected budget $B=1,000$ permits $n(1+10s) \leq B$. Count Y as assay-accepted candidates, with $E[Y]=nsv$.

The toy model compares two screens

Screen	Proposals	Assays	Expected yield
A	333.3	66.7	6.67
B	500	50	15

Screen A: $s=0.20$, $v=0.10$ gives expected yield about 6.67 using a continuous-budget approximation. Screen B: $s=0.10$, $v=0.30$ gives 15. Screen B wins through better enrichment, even though fewer proposals pass it.

Validity degradation reverses the screen preference

Screen (s,v)	Assay cost 5	Assay cost 10	Assay cost 20
A (.20,.10)	10.00	6.67	4.00
B (.10,.30)	20.00	15.00	10.00
B degraded (.10,.10)	6.67	5.00	3.33

Raising assay cost alone leaves B ahead at every displayed cost. Reducing B's conditional acceptance v from .30 to .10 reverses the preference. Endpoint: the same one-assay acceptance.

A hard cap requires a decision rule

$$E[\min(X, 20)] = \sum_{x=0}^n \min(x, 20) P(X=x)$$

One-assay endpoint: reserve 20 assays at cost 10 each; generate $n \leq 800$ proposals at cost 1. Screen passes $X \sim \text{Binomial}(n, s)$. Separate confirmation needs another reserve.

Assay a random subset of $\min(X, 20)$ passing candidates. Then cost never exceeds $n + 200 \leq 1,000$, and $E[Y] = v E[\min(X, 20)]$.

Quantify underfill and discarded candidates

Proposals	P(underfill)	E[assays]	E[assay accepted]
200	0.466	18.315	5.494
300	0.017	19.963	5.989
800	$<1.7 \times 10^{-10}$	≈ 20.000	≈ 6.000

Same one-assay endpoint, screen B ($s=.10, v=.30$). At $n=200$ the 20 slots often underfill. At $n=800$ proposal cost plus assay reserve reaches the hard budget of 1,000.

A deadline limits the usable assay reserve

$$E[Y_D] = v E[\min(X, c(D))]; \quad c(D) = \min(20, 5 \lfloor D \rfloor)$$

Suppose all proposals are ready at day 0, one instrument completes five assays per day, and no result triggers a new proposal. By day D , capacity is $c(D) = \min(20, 5 \lfloor D \rfloor)$.

For $n=300$, $s=.10$ and $v=.30$: one day gives expected yield ≈ 1.50 ; four days give ≈ 5.989 . Faster proposal generation cannot fill more physical assay slots.

One round of validation changes the next campaign

Fixed-policy pilot: 200 proposals yield 20 screen passes and 6 acceptances by the prespecified assay. Estimates are $s=.10$ and $v=.30$.

The approximate 95% Wilson interval for v is $[.145, .519]$. With 20 filled future slots and stable rates, the implied mean-yield interval is $[2.91, 10.38]$.

This is uncertainty in the mean, not the future count. A new selection policy or assay requires recalibration.

Extension: a second confirmation changes the endpoint

$$E[Y_{\text{confirm}}] = B s v q / [1 + s(10 + v c_c)]$$

Now change the endpoint: each assay-positive receives a second test at cost c_c and passes it with conditional probability q . This is an expected-budget extension.

For screen B, $c_c=20$ and $q=.8$ give about 11.54 first-assay positives and 9.23 confirmations. The earlier 15 positives used a different budget allocation.

The model fails if the screen exploits its own proxy

If the generator adapts to screen artifacts, the conditional validation rate v can fall with search effort.

Then fixed- v extrapolation overstates yield. Independent validation is required to estimate the changing selection distribution.

The same formal shell has different parameters

A cap-set checker may have near-deterministic validity but costly optimality proof. A program evaluator may have incomplete tests. Molecular and mechanism tasks add experimental uncertainty.

The common shell organizes the comparison; domain evidence supplies the parameters and scope.

A capability score cannot determine scientific value alone

Two systems with equal predictive accuracy can differ in exploration, feedback use, validation and access to instruments.

Even equal validated yield can differ in novelty, usefulness and what later discoveries it enables.

A held-out expansion test separates two yield models

Model	20 slots: calibration	40 slots: prediction
Q: quality preserved	$20 \times .30 = 6$	$40 \times .30 = 12$
D: broader policy dilutes quality	$20 \times .30 = 6$	$40 \times .15 = 6$
Observed $k=12$	Not used for fitting	Likelihood ratio ≈ 17.84 for Q

At 40 slots, use a predeclared broader selection policy. Both models assume iid outcomes within that policy. This tests the joint expansion, not assay capacity alone.

Prediction 1: evidence limits survive faster computation

Compare two simulated mechanisms with identical allowed observation laws. Vary compute while fixing data, target and prior information.

Prediction: expected discrimination stays at chance under equal priors. Reproducible above-chance performance on held-out data challenges the claimed observational equivalence.

Prediction 2: the critical path limits fixed-process speedup

Use the L05 four-operation workflow: $W=14$ h and critical path $D=11$ h. Increase workers while keeping operation times and dependencies fixed.

Prediction: completion never precedes 11 h. A measured completion below 11 h falsifies the stated process model or its timing assumptions.

Prediction 3: redundant samples give weak coverage gains

Freeze a generator and the acceptance rule. Compare independent proposals with exact duplicated proposals at equal evaluated count.

Prediction: independent success follows $1-(1-p)^n$; duplicates do not add coverage. A systematic departure in repeated trials rejects the iid fixed-p model.

Prediction 4: the bottleneck can move after improvement

Randomize proposal budget at fixed assay capacity and a frozen screen. Measure passing candidates, assayed candidates and independent confirmations.

Prediction: once assays fill, more proposals help only through better selection. Rising yield with unchanged selection quality would reject the fixed-rate pipeline model.

Prediction 5: transfer depends on a preserved structure

For causal transport, estimate source responses within strata and independently randomize treatment in a target validation sample.

Prediction: reweighting by target stratum frequencies matches target response.
Persistent within-stratum differences reject the assumed invariant response.

Prediction 6: reuse can reduce average fixed-task cost

Compare fresh and library-assisted solutions on a fixed distribution of compatible tasks. Charge library construction, adaptation and verification to total cost.

Prediction: the 20+3T toy library beats 12T from $T=3$. Failure to cross at the predicted T rejects those cost assumptions.

A map of what is established

Ingredient	Evidence type	Boundary
Computation and information limits	Formal results	Encoding, access, distribution
Feedback and scheduling	Conditional guarantees	Utility and dependency assumptions
Resource curves	Empirical models	Fitted tasks and regimes
Cross-domain transfer	Several distinct theories	Preserved object differs
Aggregate scientific progress	Models and observational data	Measurement and causal uncertainty

Theories can be rigorous without being universal; empirical evidence can be useful without identifying a single mechanism.

2026: scientific reasoning and forecasting diverge

CUSP (Cutoff-conditioned Unseen Scientific Progress) evaluates feasibility, mechanism, solution design and timing. Its July 2026 preprint reports stronger mechanism recognition than feasibility forecasting.

Historical events with temporal information controls differ from genuinely prospective forecasts. Its unresolved “Time Capsule” questions supply no completed accuracy estimate.

A forecast needs a target event, information boundary, resolution date and scoring rule. Scientific plausibility alone cannot establish forecast accuracy.

Evidence from 2025–2026: what remains unproved

Current cases do not establish a common discovery exponent, unrestricted causal transfer or universal autonomous scientific capability.

Broad task coverage with executable evaluators still leaves costly, delayed and partially identifiable empirical settings underrepresented.

Three levels of a future theory

Level 1 predicts fixed-task performance under a fixed process. Level 2 predicts changes after modifying representation, feedback or allocation.

Level 3 predicts how accumulated knowledge changes task populations and long-run outcomes. Each level adds identification problems.

A dynamic frontier requires counterfactuals

To measure the value of a discovery, compare later work with what would have happened without the new knowledge or tool.

The counterfactual is often unobserved, adoption is selective and effects can diffuse across fields.

The final claim ledger

Claim	Status	Required warrant
Work/depth lower bound	Theorem in stated model	Dependency and cost assumptions
Allocation / pilot yields	Constructed model	Specified costs and stationarity
Published scaling curves	Empirical regularities	Observed tasks and resource range
A universal discovery law	Unestablished conjecture	Cross-domain predictive tests

The course combines established results with explicit models and empirical evidence. Their warrants remain different.

A discovery can change the next problem

The resistance case began by changing the question. A new lemma, library, assay or explanatory variable can change what later investigators can ask—not only how quickly they answer.

A predictive theory must distinguish solving today's fixed task from creating tomorrow's representation, evidence channel or task population.

Conclusion: toward conditional prediction

A predictive theory of discovery should explain which resource matters, for which task, through which mechanism, and under which validation standard.

The literature provides powerful local answers. Connecting them while preserving their assumptions is the next intellectual step.

References

Stephen A. Cook (1971). The Complexity of Theorem-Proving Procedures. STOC 1971, 151–158; conference paper

Claude E. Shannon (1948). A Mathematical Theory of Communication. Bell System Technical Journal 27, 379–423, 623–656; journal article

Judea Pearl and Elias Bareinboim (2014). External Validity: From Do-Calculus to Transportability Across Populations. Statistical Science 29, 579–595; journal article; arXiv posted 2015

Nicholas Bloom, Charles I. Jones, John Van Reenen and Michael Webb (2020). Are Ideas Getting Harder to Find?. American Economic Review 110, 1104–1144; journal article

References · continued

Dennis V. Lindley (1956). On a Measure of the Information Provided by an Experiment. Annals of Mathematical Statistics 27, 986–1005; journal article

Qian Yue Hao, Fengli Xu, Yong Li and James Evans (2026). Artificial intelligence tools expand scientists' impact but contract science's focus. Nature 649, 1237–1243; journal article

Lingfei Wu, Dashun Wang and James A. Evans (2019). Large teams develop and small teams disrupt science and technology. Nature 566, 378–382; journal article

Edwin B. Wilson (1927). Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association 22, 209–212; journal article

References · continued

[Haotian Ye et al. \(2026\). Structured Scaling of AI Discovery Across Diverse Scientific Domains. arXiv:2604.19341v2; preprint, 27 July 2026](#)

[Bernardino Romera-Paredes et al. \(2024\). Mathematical discoveries from program search with large language models. Nature 625, 468–475; online 14 December 2023](#)

[Sumit Gulwani, Oleksandr Polozov and Rishabh Singh \(2017\). Program Synthesis. Foundations and Trends in Programming Languages 4, 1–119; survey monograph](#)

[Martin Pacesa et al. \(2025\). One-shot design of functional protein binders with BindCraft. Nature 646, 483–492; online 27 August 2025](#)

References · continued

Richard P. Brent (1974). The Parallel Evaluation of General Arithmetic Expressions. Journal of the ACM 21, 201–206; journal article

Gene M. Amdahl (1967). Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities. AFIPS Spring Joint Computer Conference, 483–485; conference paper

Joseph L. Watson et al. (2023). De novo design of protein structure and function with RFdiffusion. Nature 620, 1089–1100; journal article

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira and Jennifer Wortman Vaughan (2010). A Theory of Learning from Different Domains. Machine Learning 79, 151–175; journal article

References · continued

Kevin Ellis et al. (2021). DreamCoder: Bootstrapping Inductive Program Synthesis with Wake-Sleep Library Learning. PLDI 2021; conference paper; preprint has earlier title

Benjamin F. Jones (2009). The Burden of Knowledge and the Death of the Renaissance Man: Is Innovation Getting Harder?. Review of Economic Studies 76, 283–317; journal article

Niranjan Srinivas, Andreas Krause, Sham Kakade and Matthias Seeger (2010). Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. ICML 2010, 1015–1022; conference paper

Sean Wu et al. (2026). Scientific reasoning does not reliably translate into scientific forecasting in frontier AI. arXiv:2605.22681v2; preprint, 18 July 2026

References · continued

Thomas Hubert et al. (2025). Olympiad-level formal mathematical reasoning with reinforcement learning. Nature; journal article, 2025

Mert Yuksekgonul et al. (2026). Learning to Discover at Test Time. arXiv:2601.16175v2, 5 February 2026; preprint

Benjamin Fry, Kaia Slaw and Nicholas F. Polizzi (2026). Zero-shot design of drug-binding proteins via neural iterative selection–expansion. Nature 656, 237–249; online 24 June 2026

Ramon Viñas Torné et al. (2025/2026). Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. Nature Biotechnology 44, 1050–1059 (2026); online 2025

References · continued

[Alexander Novikov et al. \(2025\). AlphaEvolve: A coding agent for scientific and algorithmic discovery. arXiv:2506.13131; white paper/preprint](#)

[Juraj Gottweis et al. \(2026\). Accelerating scientific discovery with Co-Scientist. Nature 655, 487–496; journal article](#)

[Martin L. Weitzman \(1998\). Recombinant Growth. Quarterly Journal of Economics 113, 331–360; journal article](#)