LLMAutoSciLab: Closed-Loop Scientific Discovery via Active Experimentation with Large Language Models

Sanchit Kabra*, Nikhil Abhyankar*, Saaketh Desai, Prasad Iyer, Chandan K. Reddy

* Equal Contribution







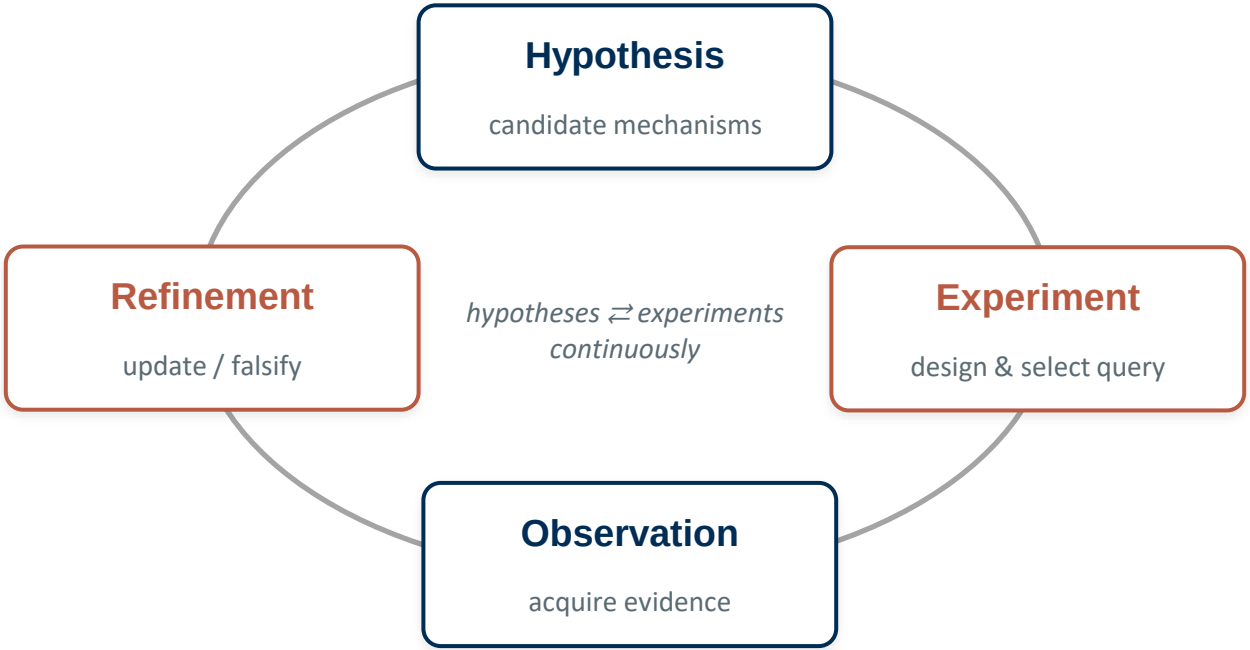
VIRGINIA TECH

Sandia National Laboratories

Real Scientific Discovery Is Closed Loop

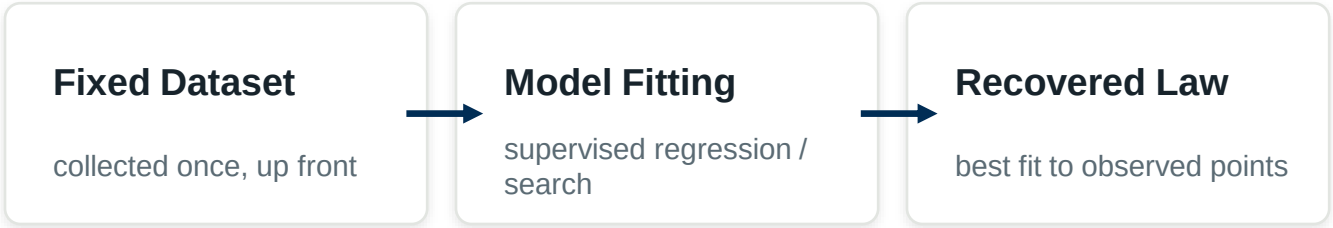
Is Scientific Discovery Static?

Scientists don't fit models to static, pre-collected data. They form a hypothesis, design an experiment that will make competing theories disagree, observe the outcome, and revise, iteration after iteration.



Existing Systems for Scientific Discovery

✓ **Problem:** Existing discovery methods treat discovery as a static problem over pre-collected, fixed datasets resulting in the **identifiability gap**.



No loop back — new data is never acquired to test or separate competing hypotheses

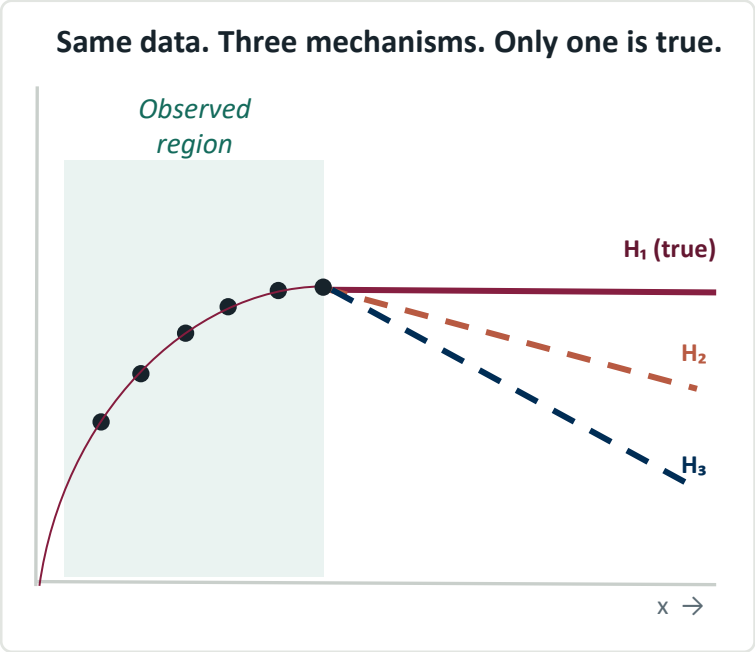
Why this falls short:

- Assumes the dataset is already fully observed and sufficient
- No mechanism to acquire new, targeted data when the fit is ambiguous
- Treats discovery as curve-fitting, not as resolving competing mechanisms

Bottleneck with Static Discovery | Identifiability Trap

The true underlying laws are hidden behind imperfect measurements and limited data.

- **Data Limits**: Data from a limited range of observations results in identifiability issues, hindering the discovery of true laws.
- **The Identifiability Trap:** With limited, passively-collected observations, many structurally different mechanisms can fit the same data equally well locally and then diverge wildly outside it.
- Static datasets simply cannot tell them apart.



Bottleneck with Static Discovery

- **Existing Methods**

- Self-driving labs and Bayesian active learning design experiments to reduce **predictive uncertainty or optimize a condition**, not to resolve which mechanism is correct.
- Active learning methods **rely on humans** for hypothesis formulation.



What's missing: Automated active experimentation based on hypothesis disambiguation to identify the true law.

- **Existing Benchmarks**

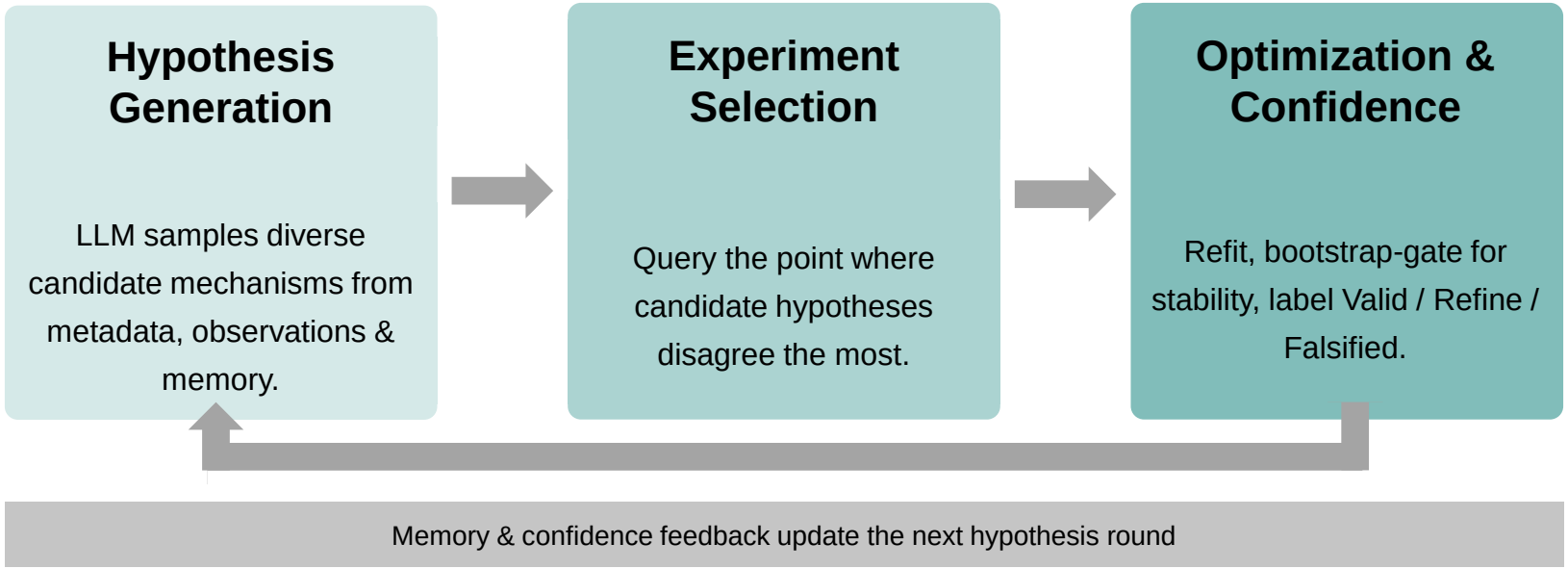
- Benchmarks built to test scientific discovery hand the learner **a fixed dataset**.
- Most existing benchmarks **limit scientific discovery to symbolic discovery** and are **prone to memorization** or **based on counterfactual laws defying real-world physical laws**.



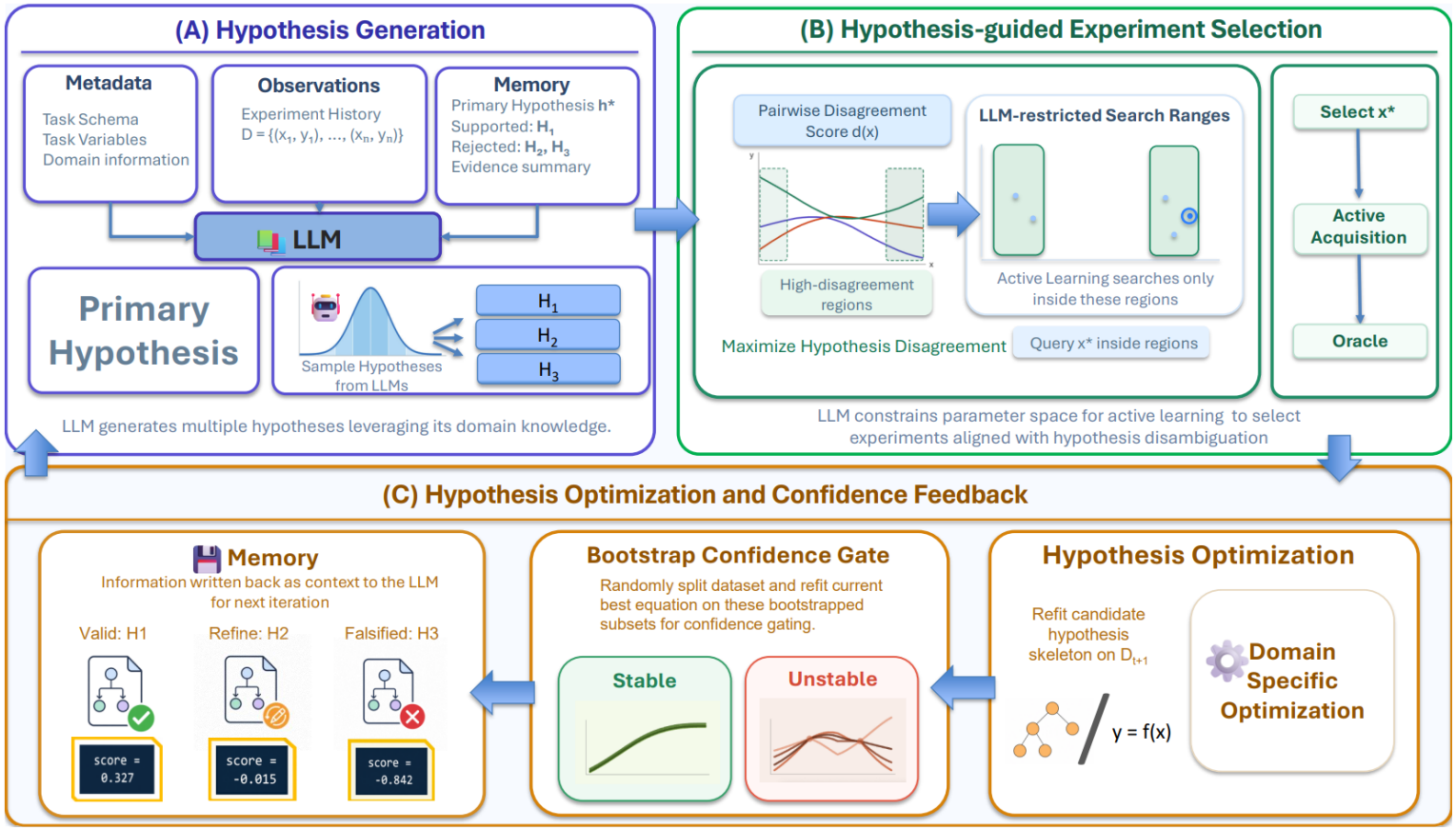
What's missing: A physically grounded benchmark that combines active exploration with open-ended discovery beyond symbolic regression in one benchmark.

Our Approach

Instead of fitting one final model, the system continuously proposes candidate mechanisms, hunts for the experiment that best separates them, and updates its beliefs from the result.



LLMAutoSciLab | Framework



LLMAutoSciLab | Hypothesis Generation

The LLMAutoSciLab generates multiple hypotheses leveraging:

- **Task Metadata**

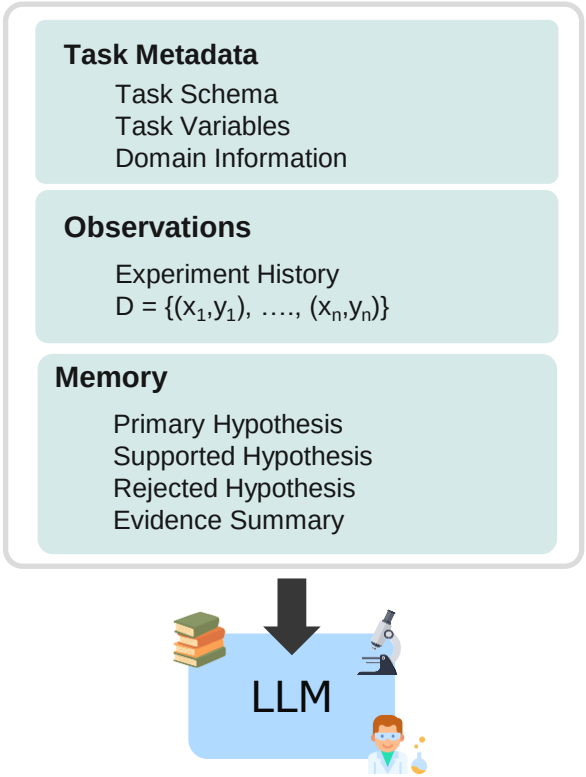
Acts as the foundational domain schema. It provides structural definitions containing the superset of physical variables, parameters, and experimental limits defining the environment constraints.

- **Observations**

Feeds previous experimental observations D directly to the LLM. Ingesting successive trial inputs and metrics keeps the search space grounded in factual, physical data.

- **Memory**

Tracks dynamic hypotheses (primary, supported, and rejected) alongside structural summaries of current evidence, assisting the model in tracking discovery context iteratively.



LLMAutoSciLab | Hypothesis Generation

To avoid premature search collapse and promote hypothesis diversity, we employ:

✓ Two-model decoupling

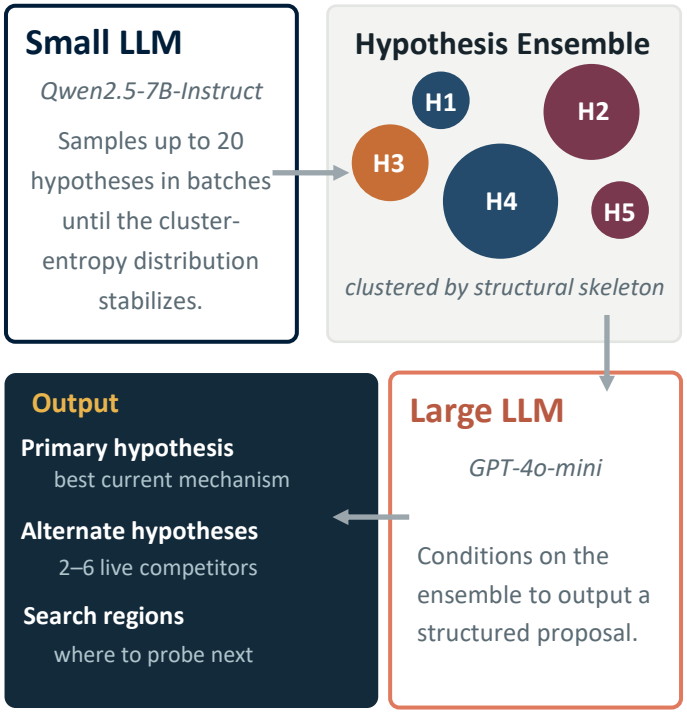
- A smaller LLM proposes diverse candidate hypotheses conditioned on the current state.
- A larger LLM synthesizes these into a structured proposal a primary hypothesis plus alternatives — avoiding premature collapse to a single guess from sparse data.

✓ Diversity over premature commitment

- Rather than outputting one final answer, the LLM defines a *hypothesis space* consisting of multiple structurally distinct, data-plausible mechanisms.

✓ Search-region proposal

- The large LLM also outputs search regions in parameter space where the competing hypotheses are expected to be most distinguishable, setting up the next stage.



LLMAutoSciLab | Hypothesis Guided Experiment Selection

✓ **LLM-Restricted Ranges**

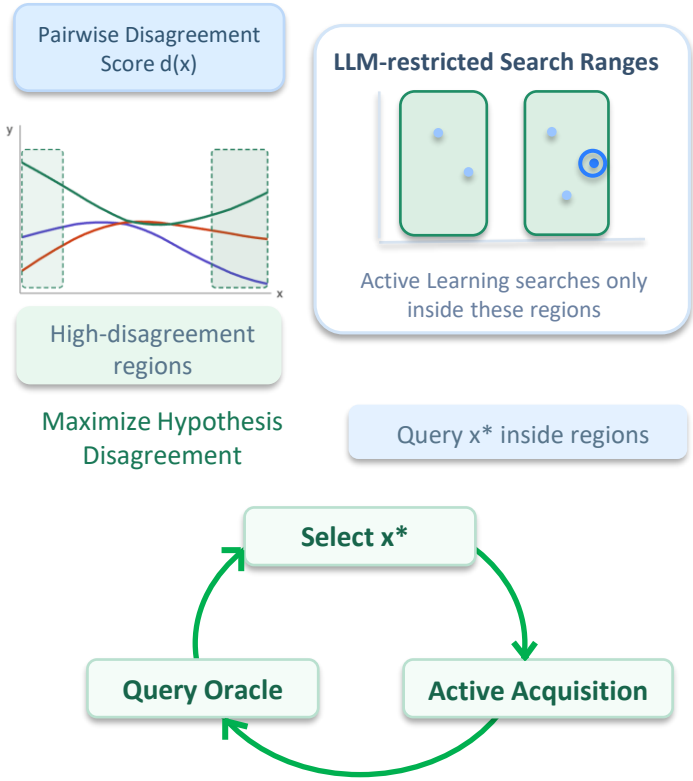
- Instead of searching the entire parameter space, we restrict queries solely to critical high-disagreement boundary regions proposed by the LLM.
- This narrows focus to where predictions diverge.

✓ **Maximizing Disagreement**

- The system evaluates pairwise disagreement scores across competing hypotheses.
- By targeting the exact point x* with maximum prediction variance, it maximizes the information gain of each trial.

✓ **Active Data Acquisition**

- Once selected, the point x* is dispatched to query the physical oracle.
- The resulting observation (x*,y*) is appended to update the dataset, continually refining prediction accuracy.



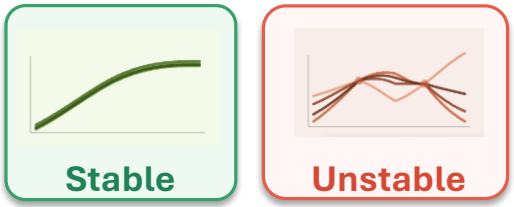
LLMAutoSciLab | Hypothesis Optimization and Feedback

Hypothesis Optimization

Refits candidate hypothesis skeletons on the updated dataset D_{t+1} using domain-specific continuous parameter optimization to produce accurate rate equations.

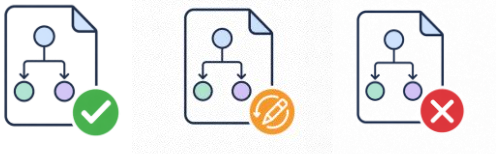
Bootstrap Confidence Gate

Randomly splits datasets into bootstrapped subsets and refits formulas to determine equation stability, isolating robust paths from unstable fits.



Memory

Categorizes optimized equations with computed confidence scores, sorting them into active memory as context for the next LLM loop iteration.



ActiveSciBench | Active Scientific Discovery Benchmark

ActiveSciBench-Chem

- **Objective:** Reconstruct precise symbolic rate laws and identify active parameters from sequential physical trials.
- **Interface:** Shared 7-dimensional assay interface containing input concentrations, temperature, loading, and pH levels.
- **Construction:** 57 curated tasks spanning three progressive difficulty tiers:
  - **Easy:** Covers fundamental biological rate laws
  - **Medium:** Structural mixtures
  - **Hard:** Non-linear dynamics

Metrics
· Symbolic Accuracy · Exact Accuracy · RMSLE

ActiveSciBench-GRN

- **Objective:** Recover complete signed, directed regulatory graphs and trace their non-linear cellular dynamics.
- **Interface:** Discretized node environments allowing directed interventions (knock up, knock down, or targeted node perturbation).
- **Construction:** 45 tasks mapped over 5 canonical motif structures:
  - Incoherent feedforward loop
  - Negative-feedback circuit
  - Toggle-switch / bistable circuit
  - Activation chain
  - Coherent feedforward loop

Metrics
· Edge F1 · Exact Graph Accuracy · Sign Accuracy

ActiveSciBench | Benchmarking Active Scientific Discovery

STATIC BENCHMARKs

Dataset handed over

↓

Fit a curve or recall a law

↓

An equation

✗ memory or curve fitting can score

ACTIVESCIBENCH: A SANDBOX FOR DISCOVERY

THE AGENT

Input: The variables and domain knowledge

Starts with no data

Loops: hypothesize → experiment → update beliefs

experiment x →

each one costs budget

← **data y**

THE SANDBOX

Hides: a real law or gene graph

Takes: an assay or a gene intervention

Returns: data only for what is run

Output: The law or graph, and which inputs matter. ✓ **Recall cannot score here as data needs to be found.**

WHAT IS NEW COMPARED WITH EXISTING BENCHMARKS

1 Real, hidden mechanisms

✗ **Existing:** Made-up, counterfactual physics

✓ **ActiveSciBench:** Real enzyme kinetics and gene regulation; each instance hidden

2 Beyond equations

✗ **Existing:** Only symbolic expressions are scored

✓ **ActiveSciBench:** Symbolic rate laws and signed causal gene graphs

3 Inputs must be discovered

✗ **Existing:** Relevant variables handed over

✓ **ActiveSciBench:** Distractor inputs sit among the ones that matter

4 Recall is not enough

✗ **Existing:** Data handed over, so LLM memory or fitting can work

✓ **ActiveSciBench:** No data at the start; every observation costs an experiment, so recovery is related to experiment selection and data fitting

Discovery is not data fitting: choosing which experiment to run, the most expensive step, is the task.

ActiveSciBench-Chem | Enzyme Kinetics

TYPICAL BENCHMARK
Variables and data are given; the agent only fits a law

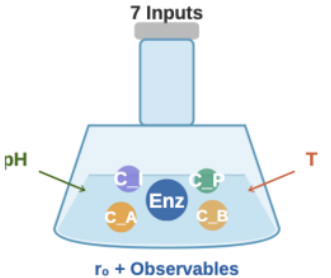
→

ACTIVESCIBENCH-CHEM
The agent must find which inputs matter and generate its own data, then recover the law

EXPERIMENTAL SYSTEM

Reaction: E + S → P

Enzyme Kinetics: Rate r_0 depends on substrate, enzyme, inhibition and conditions

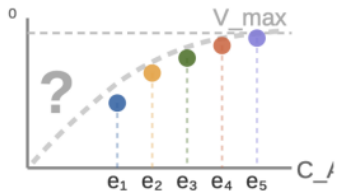


r_0 + Observables

Not all inputs appear in the true law

HIDDEN ORACLE

True Rate Law — Withheld



Each experiment adds one observation to fit the equation

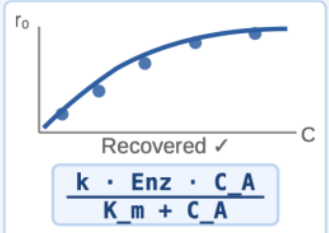
Example Experiments

- Vary C_A, fix others
- Vary T, fix C_A
- High C_A + High Enz

Budget B Experiments

TARGET OUTPUT

Symbolic Rate Law



✓ C_A, Enz ✗ C_B, C_I, C_P, T, pH

Metrics

- ✓ Symbolic Accuracy
- ✓ Numerical Exact Acc.
- ✓ RMSLE

WHAT THE VARIABLES MEAN

r_0	reaction rate: all the sandbox reports
C_A	substrate A concentration
C_B	second substrate concentration
C_I	inhibitor concentration
C_P	product concentration
Enz	enzyme loading (how much enzyme)
T, pH	temperature and acidity
k, K_m	hidden constants: k·Enz = V_{max}, the top rate; K_m = C_A at half of it

In this task the law uses only **C_A** and **Enz**; the other five inputs are distractors the agent must rule out.

No data at the start
r_0 is revealed only for the assays the agent chooses to run

Find what matters
7 inputs, but only C_A and Enz appear; the rest must be ruled out

Real, but hidden
textbook families, yet this instance's form and constants are unknown

Every assay costs
a fixed budget, so choosing informative assays is the discovery

ActiveSciBench-GRN | Gene Regulatory Networks

TYPICAL BENCHMARK
Benchmarks focus only on equation discovery

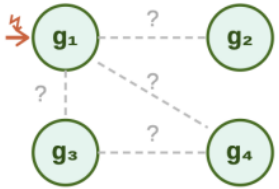
→

ACTIVESCIBENCH-GRN
Goes beyond equations: the agent must find the right signed causal graph

SYSTEM NODES

Intervenable Nodes

- $g_1 \ldots g_n$ gene expression
- σ_i upstream signal
- k_i kinetic param.



Perturb any node,
observe all downstream Δ

→

HIDDEN ORACLE

True Graph G + Dynamics

? ? ?
Edges + Signs Unknown

Each experiment adds one observation to fit the graph

Example Experiments

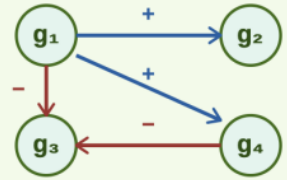
- Knock up g_1 → cascade
- Knock down g_2 → repressors
- Perturb g_1+g_3 → break ties

Budget B Experiments

→

TARGET OUTPUT

Signed Directed Graph



+ Activates − Represses

Metrics

- ✓ Exact Graph Accuracy
- ✓ Edge F1 Score
- ✓ Sign Accuracy

WHAT THE VARIABLES MEAN

$g_1 \cdots g_n$	expression levels (what is measured)
σ_i	upstream signal that drives a gene
k_i	hidden kinetic parameters (rates)
+ edge	activates: target follows the source
− edge	represses: target moves against it
up / down	knock a gene's expression up or down
perturb	nudge one or more genes at once

In this task **g_1** activates g_2 and g_4 and represses g_3; **g_4** also represses g_3. The agent must find all four edges and their signs.

No data at the start
Expression changes appear only after the agent intervenes

Correlation misleads
Genes that move together may be linked only indirectly; knockdowns tell

Signed and directed
The answer is a causal graph with activation and repression, not an equation

Every intervention costs
A fixed budget; multi-node perturbations break ties between rival graphs

LLMAutoSciLab | Quantitative Results

Method	ED	Easy			Medium			Hard			Overall		
		SA(↑)	Ex.(↑)	RMSLE(↓)	SA(↑)	Ex.(↑)	RMSLE(↓)	SA(↑)	Ex.(↑)	RMSLE(↓)	SA(↑)	Ex.(↑)	RMSLE(↓)
NewtonBench													
PySR	×	38.89	88.89	0.030	27.78	75.00	0.125	5.56	59.72	**0.398**	24.07	74.54	0.182
Bayesian Optimization	✓	40.28	84.72	0.070	27.78	70.83	0.212	5.56	50.00	0.730	24.54	68.52	0.334
Bayesian Experimental Design	✓	19.44	78.57	0.307	11.11	66.18	0.449	2.78	45.31	1.010	11.11	63.86	0.577
LLM-only	✓	16.67	19.44	3.572	2.78	2.78	4.738	0.00	0.00	6.757	6.48	7.41	5.039
Code-assisted LLM	✓	13.89	19.44	4.372	5.56	11.11	5.437	2.78	0.00	4.897	7.41	10.19	4.912
LLM-AutoSciLab (Ours)	✓	**79.20**	**93.10**	**0.018**	**72.20**	**84.70**	**0.040**	**51.40**	**66.70**	0.404	**67.60**	**81.50**	**0.150**
ActiveSciBench-Chem													
PySR	×	33.33	33.33	0.455	0.00	11.54	1.545	0.00	0.00	0.601	7.89	15.79	1.212
Bayesian Optimization	✓	22.22	14.29	2.001	0.00	3.85	1.338	0.00	0.00	6.564	5.26	6.98	1.960
Bayesian Experimental Design	✓	**77.78**	77.78	**0.028**	19.23	30.77	0.333	0.00	0.00	0.187	31.58	39.47	0.249
LLM-only	✓	0.00	0.00	0.563	0.00	0.00	0.777	0.00	0.00	0.849	0.00	0.00	0.764
Code-assisted LLM	✓	0.00	0.00	0.554	0.00	0.00	0.869	0.00	0.00	0.894	0.00	0.00	0.824
LLM-AutoSciLab (Ours)	✓	55.56	**88.89**	0.298	**22.22**	**37.04**	**0.179**	**42.86**	**52.38**	**0.154**	**35.09**	**50.88**	**0.189**
		F1(↑)	**Ex.(↑)**	**Sign(↑)**	**F1(↑)**	**Ex.(↑)**	**Sign(↑)**	**F1(↑)**	**Ex.(↑)**	**Sign(↑)**	**F1(↑)**	**Ex.(↑)**	**Sign(↑)**
ActiveSciBench-GRN													
GENIE3	×	41.54	0.00	85.33	35.00	0.00	71.89	39.27	0.00	80.00	38.60	0.00	79.07
GIES	×	**70.12**	20.00	97.78	47.07	0.00	80.56	51.61	0.00	79.00	56.27	6.67	85.78
NOTEARS	×	31.52	0.00	66.67	25.71	0.00	56.67	25.57	6.67	51.11	27.60	2.22	58.15
Random sampling	✓	33.64	0.00	86.67	33.46	0.00	86.67	39.58	6.67	80.00	35.56	2.22	84.44
Uncertainty sampling	✓	55.26	13.33	93.33	49.63	0.00	86.67	45.40	0.00	84.44	50.10	4.44	88.15
LLM-only	✓	53.61	0.00	81.67	46.11	0.00	77.78	51.51	0.00	78.89	50.41	0.00	79.44
Code-assisted LLM	✓	51.73	0.00	82.78	57.65	0.00	91.67	54.63	0.00	83.33	54.67	0.00	85.98
LLM-AutoSciLab (Ours)	✓	66.27	**26.67**	**100.0**	**83.13**	**40.00**	**100.0**	**68.06**	**26.67**	**94.44**	**72.49**	**31.11**	**98.15**

Table: Performance comparison of baselines on 63 ODEBench datasets

NewtonBench | Qualitative Example

TARGET LAW (hidden from the agent)

$$w = k/m^{1.3} - (b/2m)^{0.7}$$

restoring term − nonlinear damping

A damped oscillator with an altered law: find both terms and their exponents

Variables k, m and b are known. Below are the exact laws each method returned in the paper (Fig. 3).

correct structure wrong or spurious structure

01 NON-LLM METHODS

Fits the data, not the mechanism

PySR *spurious sine term*

$$w = (k - b)/(m(km - b + term)) + \sin(0.5\sqrt{(b \cdot m)})^{0.132}$$

Bayesian Opt. *entangled numerical fit*

$$w = (k - b)/(m \cdot (k \cdot m)) + \sin(0.535 \cdot \sqrt{(b \cdot m)}^{0.132})$$

Bayesian Exp. Design *right parts, wrong form*

$$w = \sqrt{(k^{0.868}/m^{1.130} - 0.256 \cdot b \cdot m - 0.274/k^{0.817})}$$

Each fits the observations, but adds a sine term or entangles the parts; the additive mechanism is lost.

02 LLM-ONLY / TOOL-USE

Recalls the textbook answer

LLM-only *textbook harmonic form*

$$w = \sqrt{(k/m)}$$

Code-assisted LLM *harmonic part only*

$$w = \sqrt{(k/m)} - C_0$$

A broad restoring prior, but the altered exponents and the nonlinear damping term are never tested.

03 LLM-AutoSciLab

Tests which law is true

LLM-AutoSciLab *exact structure*

$$w = k/m^{1.3} - (0.4915 \cdot b/m)^{0.693}$$

$0.4915 \approx 1/2$, so $b/2m$ $0.693 \approx 0.7$

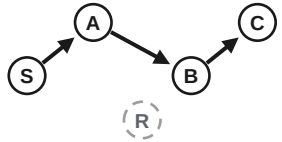
how (illustrative): rivals with and without damping disagree most at large b, so it tests there

Keeps rival laws, probes where they disagree, then fits the surviving structure.

The key is to test competing explanations, not just fit the points or recall a formula.

ActiveSciBench-GRN | Qualitative Example

GROUND TRUTH



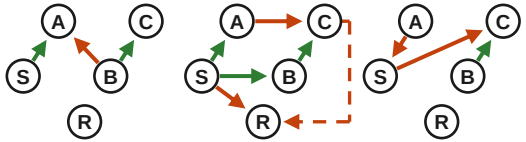
Hidden network: an activation chain S → A → B → C; gene R affects nothing

Every method returns a signed graph. Below are the actual graphs each method returned in the paper (Fig. 8).

correct edge incorrect edge spurious edge

01 NON-LLM METHODS

Confuses links and causes



GIES
A–B reversed

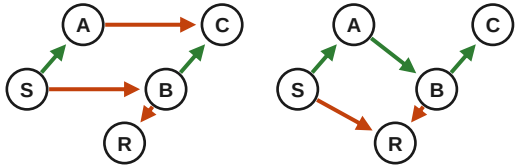
GENIE3
over-connected

NOTEARS
reverses the route

They reverse arrows, miss edges or over-connect the graph with extra and spurious edges.

02 LLM-ONLY / TOOL-USE

Adds plausible but false links



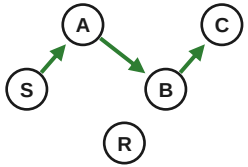
LLM-only
shortcut + extra edges

Code-assisted
chain + spurious branch

Plausible wiring, but with shortcuts and branches that no experiment supports.

03 LLM-AutoSciLab

Intervenes to test each link



LLM-AutoSciLab
exact recovery

how (illustrative)

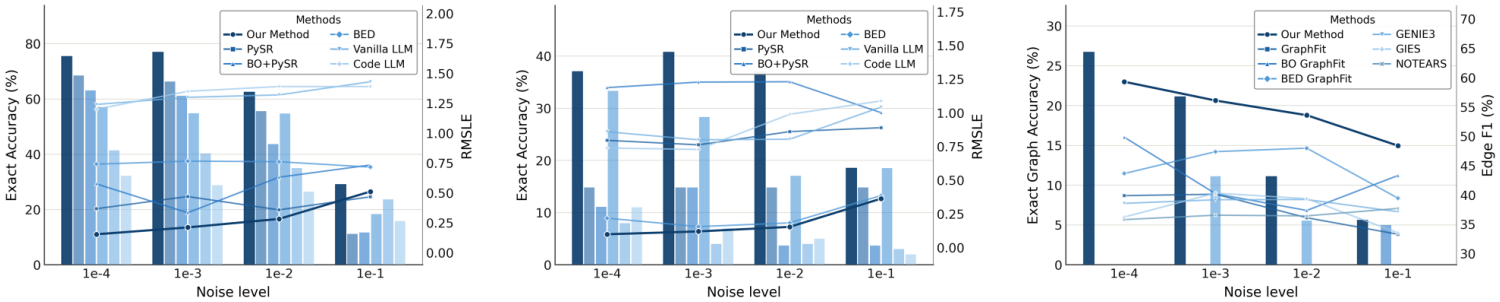
Knock down A
B falls ⇒ keep A → B

Knock up R
nothing moves ⇒ drop R

Its perturbation choices “isolate the causal backbone”: the chain is exact and R is excluded.

Changing one gene can reveal cause and direction that passive observation cannot.

LLMAutoSciLab | Analysis



(a) NewtonBench (b) ActiveSciBench-Chem (c) ActiveSciBench-GRN

Figure: Ablation study across all benchmarks removing one component from LLM-AutoSciLab.

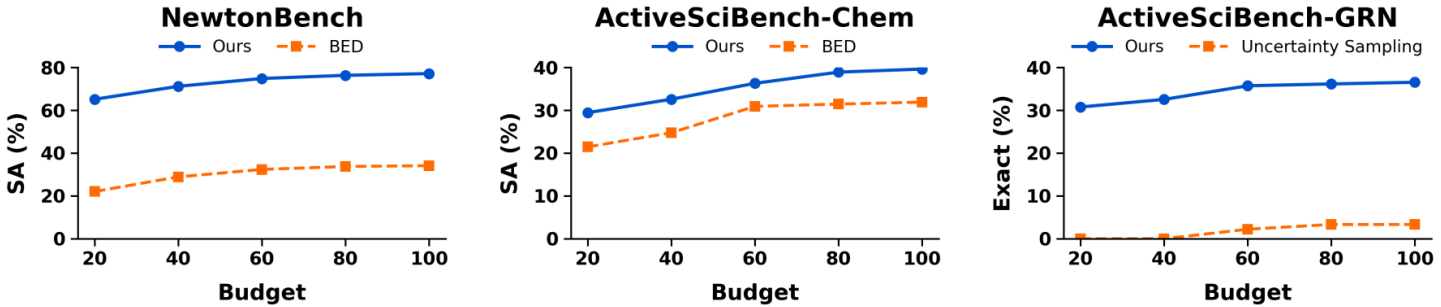


Figure: Budget ablations across benchmarks showing the recovery metric versus query budget.

LLMAutoSciLab | Analysis

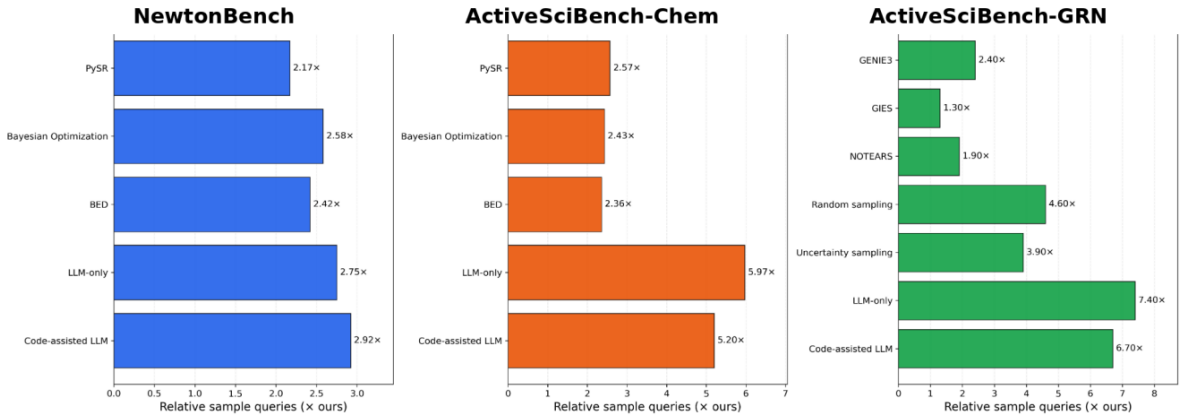


Figure: Relative sample efficiency across benchmarks. Each bar shows the multiplicative number of samples required by a comparison method to match the fixed-budget performance of LLM-AutoSciLab (lower is better).

	Newton			Chem			GRN		
Backbone	**SA**	**Ex.**	**RMSLE**	**SA**	**Ex.**	**RMSLE**	**F1**	**Ex.**	**Sign**
GPT-4o-mini	**67.60**	81.50	0.150	**35.09**	50.88	0.189	**72.49**	**31.11**	98.15
Qwen3-4B	57.78	82.86	0.174	12.92	53.56	0.3766	49.56	21.78	93.52
Qwen3-14B	54.29	86.11	0.127	23.88	50.59	0.1574	51.19	22.34	92.16
Qwen3-32B	60.56	**88.12**	**0.101**	25.56	**52.81**	**0.0815**	62.70	25.88	**98.70**

Figure: LLM backbone comparison. GPT-4o-mini is the default backbone; Qwen3 models evaluate open-weight scaling.

Summary

- LLMAutoSciLab is a closed-loop framework coupling LLM-guided hypothesis generation, hypothesis-conditioned experiment design, and bootstrap-gated refinement.
- We introduce ActiveSciBench: two new scientific discovery benchmarks with hidden relevant variables and budget-limited oracle access, beyond static function fitting.
- A disagreement-based acquisition objective that outperforms Bayesian and predictive-uncertainty methods across three benchmark families.
- We achieve state of the art results: 67.6% SA (NewtonBench), 35.1% SA (Chem), 31.1% exact graph recovery (GRN) with 2–5× better sample efficiency.

Future Directions

- Scaling acquisition to surrogate models under tighter query budgets and extending closed-loop discovery beyond ODEs to PDEs and coupled multi-physics systems.
- Expanding such dynamic discovery paradigms to more AI4Science systems to enable more interpretable, data-efficient and reliable scientific discovery.

Thank You