Towards Scientific Discovery with Generative AI: Progress, Opportunities, and Challenges

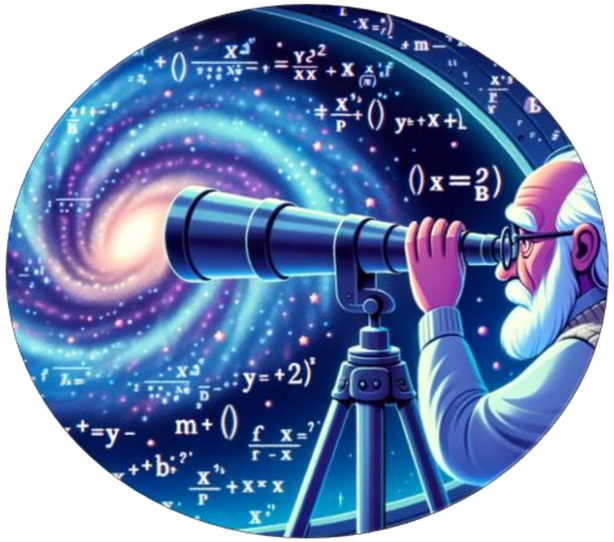
Chandan Reddy and Parshin Shojaee
Dept. of Computer Science

VIRGINIA TECH

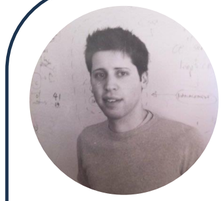
Scientific Discovery

Scientific Discovery is the process of **formulating** and **validating** **new concepts, laws, and theories** to explain natural phenomena
(Ball et al., 2020)

One of the humanity's most intellectually demanding and impactful pursuits!!



Why Scientific Discovery? Why now?



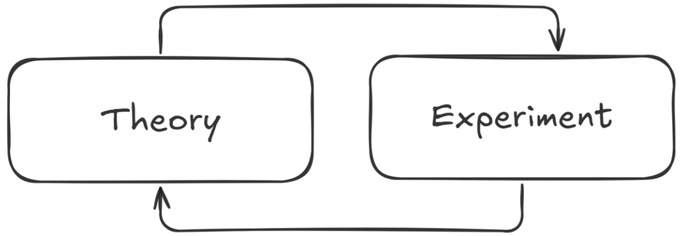
Sam Altman

Superintelligent tools could massively accelerate scientific discovery and innovation well beyond what we are capable of doing on our own, and in turn massively increase abundance and prosperity

Demis Hassabis

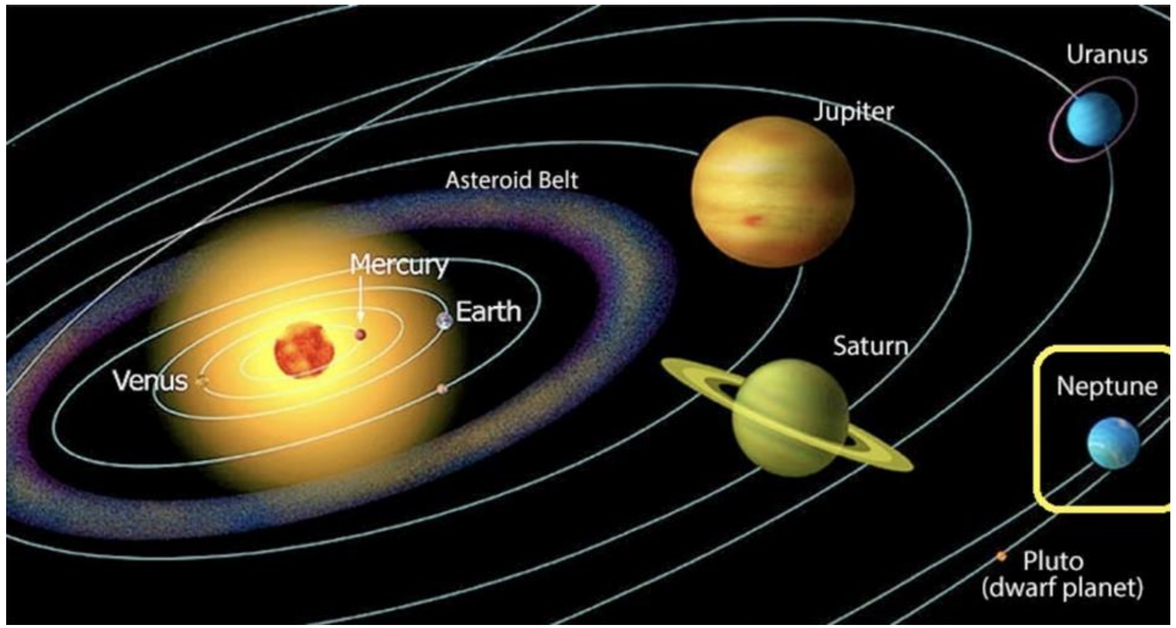
… We're on the cusp of an incredible new golden age of AI accelerated scientific discovery. Hypothesis generation and testing is a critical capability for AGI imo. …

Scientific Discovery



- **From Theory to Experiment:**
  - Uranus‘s orbit deviated from path predicted by Newton’s laws of gravity.
  - Hypothesis: there exists a planet beyond Uranus whose gravitational pull affects its orbit.
  - Discovery (1846) of “Neptune”

- **From Experiment to Theory:**
  - Mercury's orbital anomaly
  - Hypothesis: Existence of a new planet rejected!
  - Discovery (1915) of Einstein's “General Theory of Relativity”



Theory and Experiments are Complementary in Scientific Discovery!

Scientific Discovery | Why is it Challenging?

◉ Requires **deep creativity, solid reasoning, and empirical validation**.

◉ The **growing volume of domain-specific scientific knowledge** makes it sometime hard for human scientists to keep up.

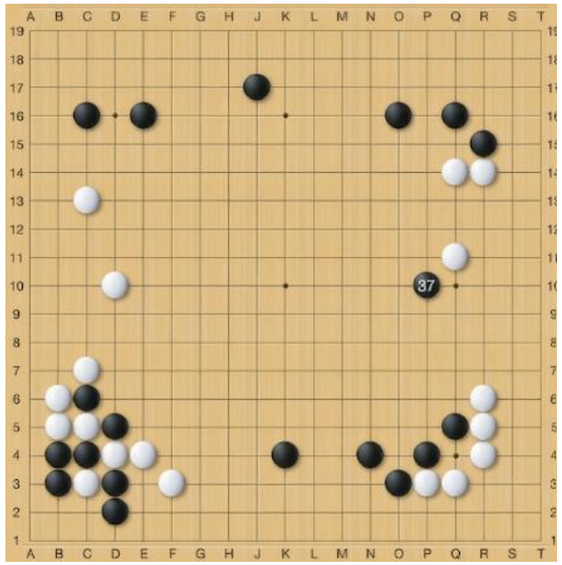
◉ Many scientific problems involve **high-dimensional, combinatorial search spaces** (e.g., equation discovery, materials discovery, drug design, protein engineering, theorem proving, …).

◉ The need for **real-world experiments and fine-grained scientific simulations** makes the process time-consuming and expensive.

AI Revolution | Combinatorial Spaces

Success in **Large Combinatorial Spaces**: **AlphaGo** Example.

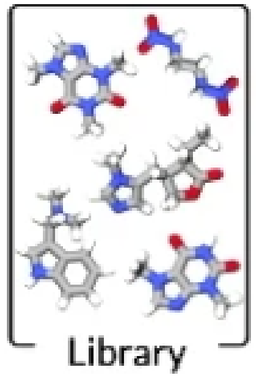
Breakthroughs in **games like Chess and Go** (AlphaGo, MuZero) prove that AI can master **complex and combinatorial search-based reasoning**.



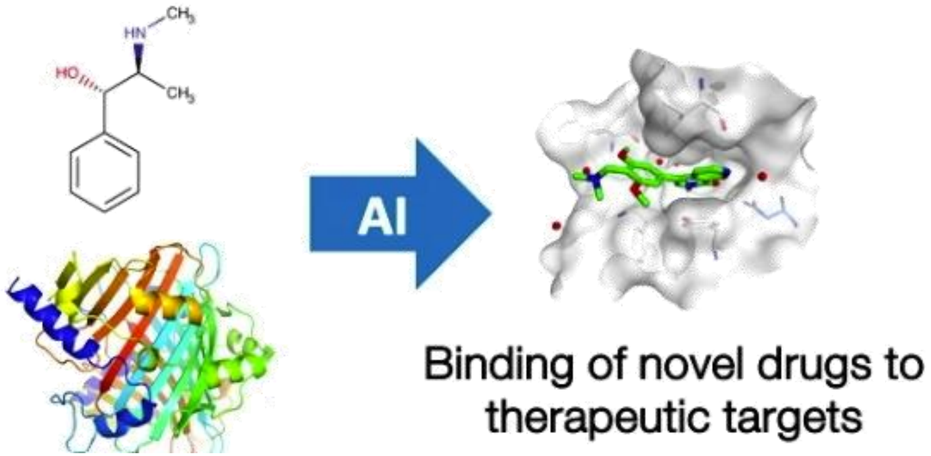
Scientific Discovery & Combinatorial Spaces

Many of the **scientific discovery problems** involve navigating **large combinatorial spaces**

- **Material Discovery**



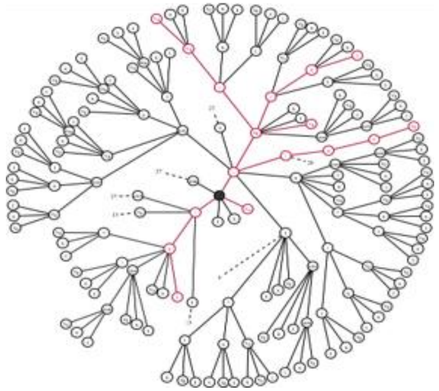
- **Drug Design**

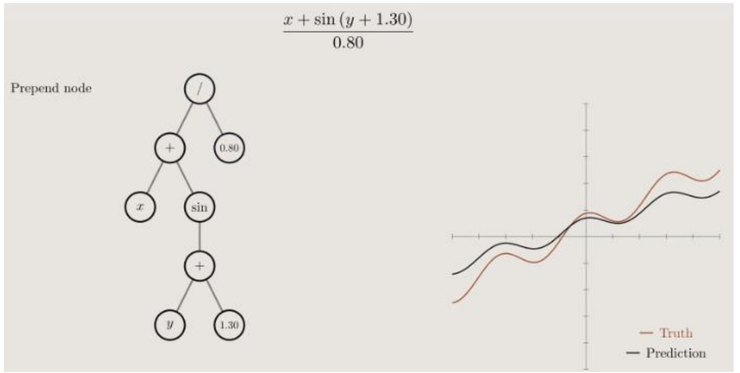


Binding of novel drugs to therapeutic targets

- **Equation Discovery**

...





Equations include both **discrete** and **continuous** components

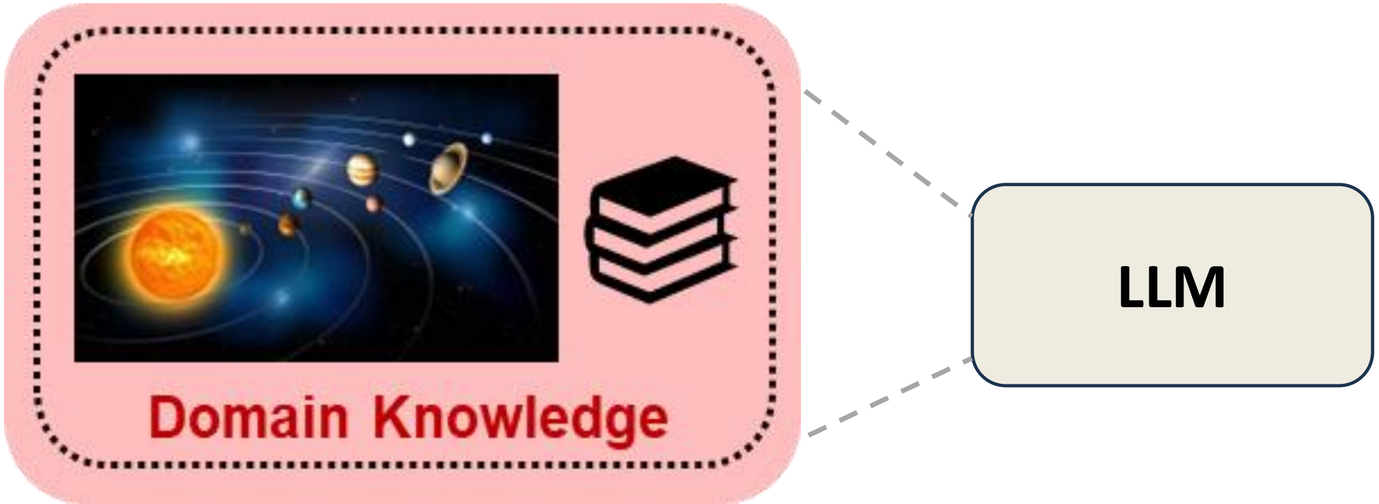
$$\frac{\mathbf{1.0}\ \boldsymbol{x} + \boldsymbol{sin}(\boldsymbol{y} + \mathbf{1.31})}{\mathbf{0.5}}$$

* Plots borrowed from (Cranmer et al., 2023) and (Zitnik et al., 2023)

AI Revolution | LLMs

LLMs and Vast Embedded Scientific Knowledge

- Large Language Models (LLMs) are trained on **vast corpora of scientific literature and books**
- They can assist **to integrate scientific domain knowledge** into the **discovery process**
- They are also showing **emergent capabilities in the reasoning and problem-solving**

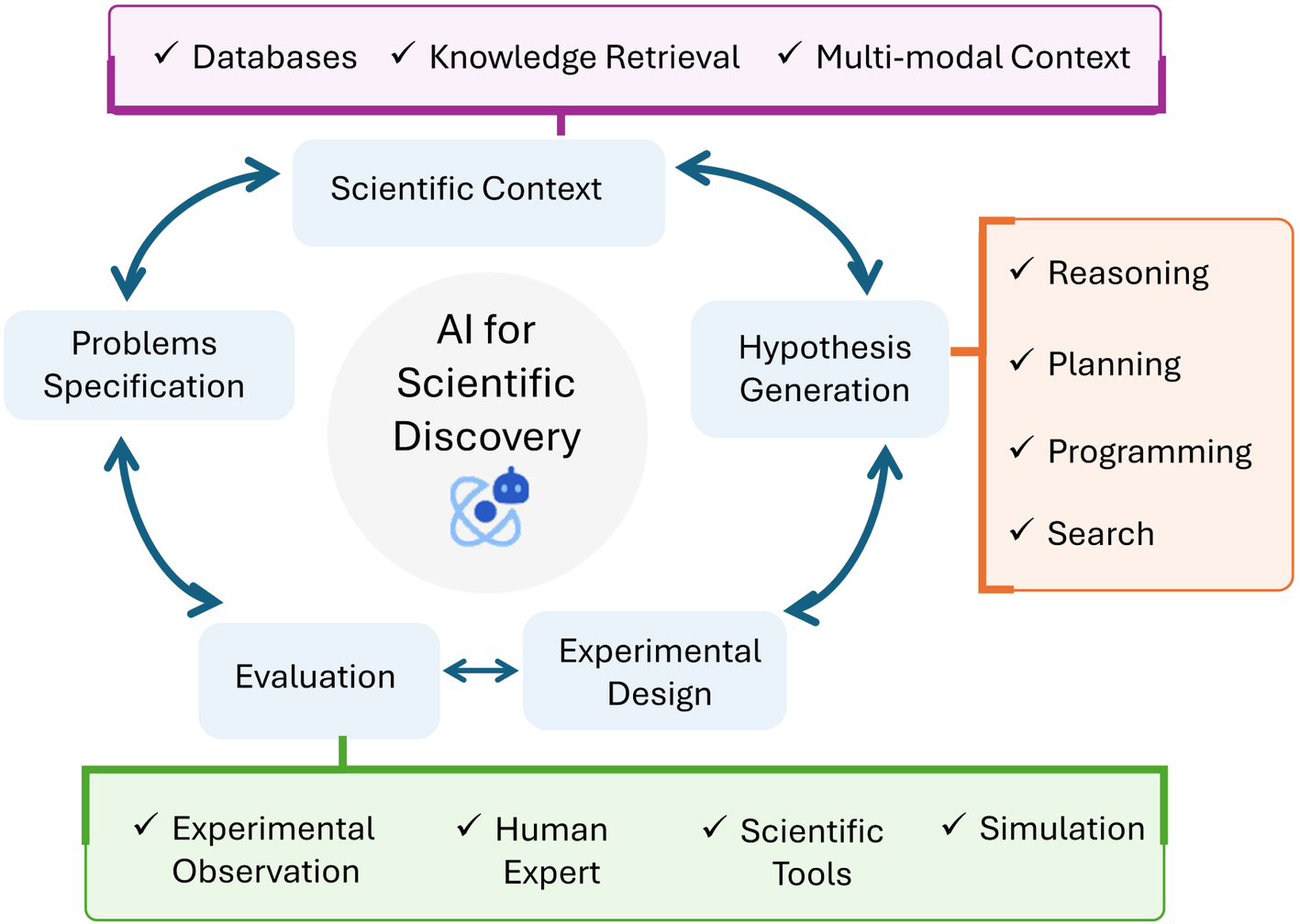


AI for Scientific Discovery (Cycle)

Iterative process of scientific inquiry

Begins with user-defined specification

- Retrieves relevant scientific context (literature & databases)
- Utilize generative AI to generate
  - new hypotheses
  - new experimental designs
- Refine AI-generated concepts with:
  - Experimental observation
  - Expert input
  - Scientific tools



Recent AI Advancements for Scientific Tasks

Literature Analysis and Brainstorming

LLMs pre-trained on vast amounts of scientific corpora can help

- Literature information retrieval
- Summarization
- Question-answering
- Brainstorming ideas

Theorem Proving

Formal theorem derivation has a fundamental role in science

- Language models trained on large-scale proofs
- Integration with proof assistants (lean, isabelle, …)
- Proof search with learned priors

Experimental Design

A critical and time-consuming process in science, requiring extensive domain knowledge and planning

- LLM agents to design, plan, and even execute experiments in simulation settings

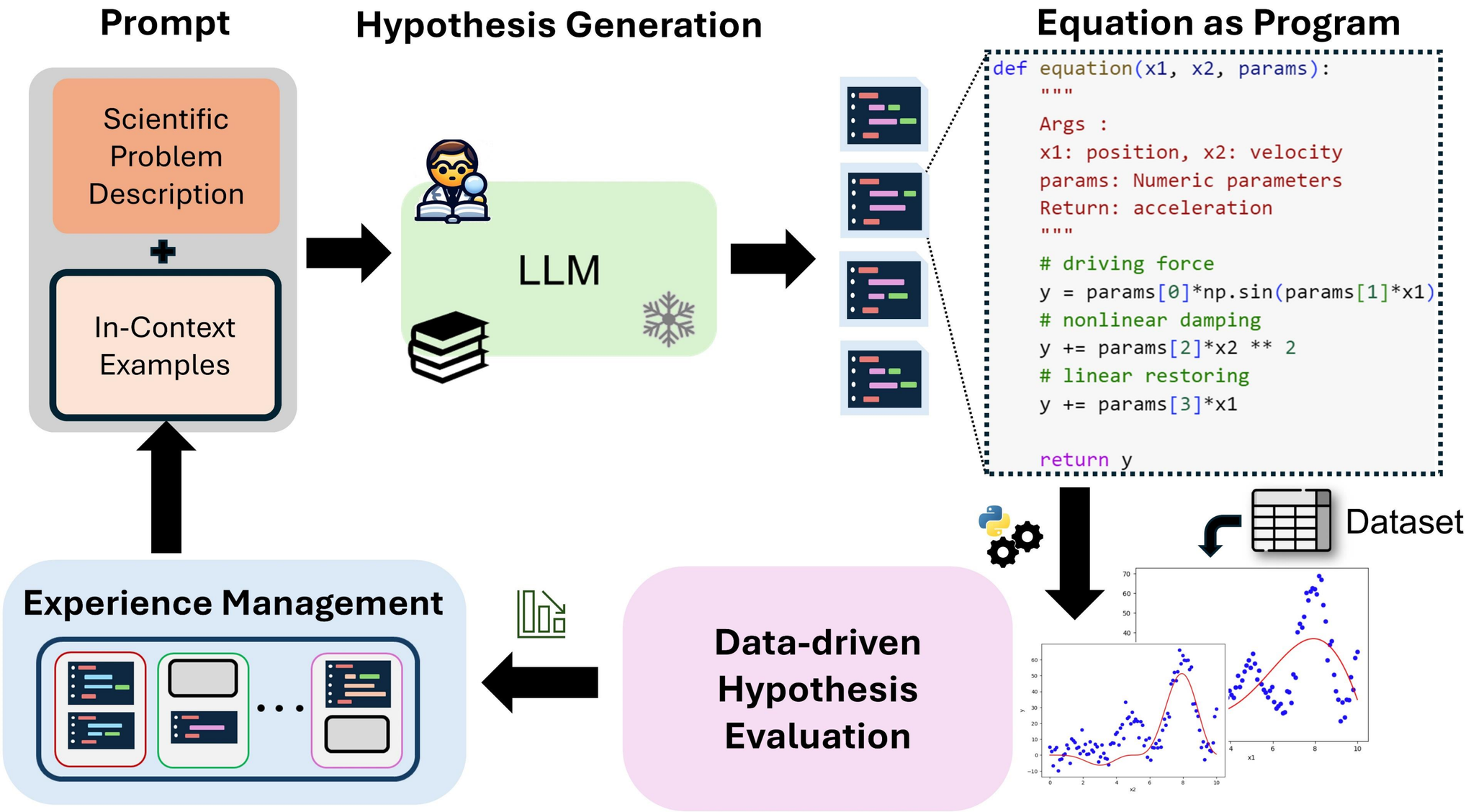
Data-driven Discovery

Leveraging ever-growing volumes of observational, and synthetic data to uncover new patterns, relationships, and laws

- Equation Discovery from data
- Finding new material structures (Material Discovery)

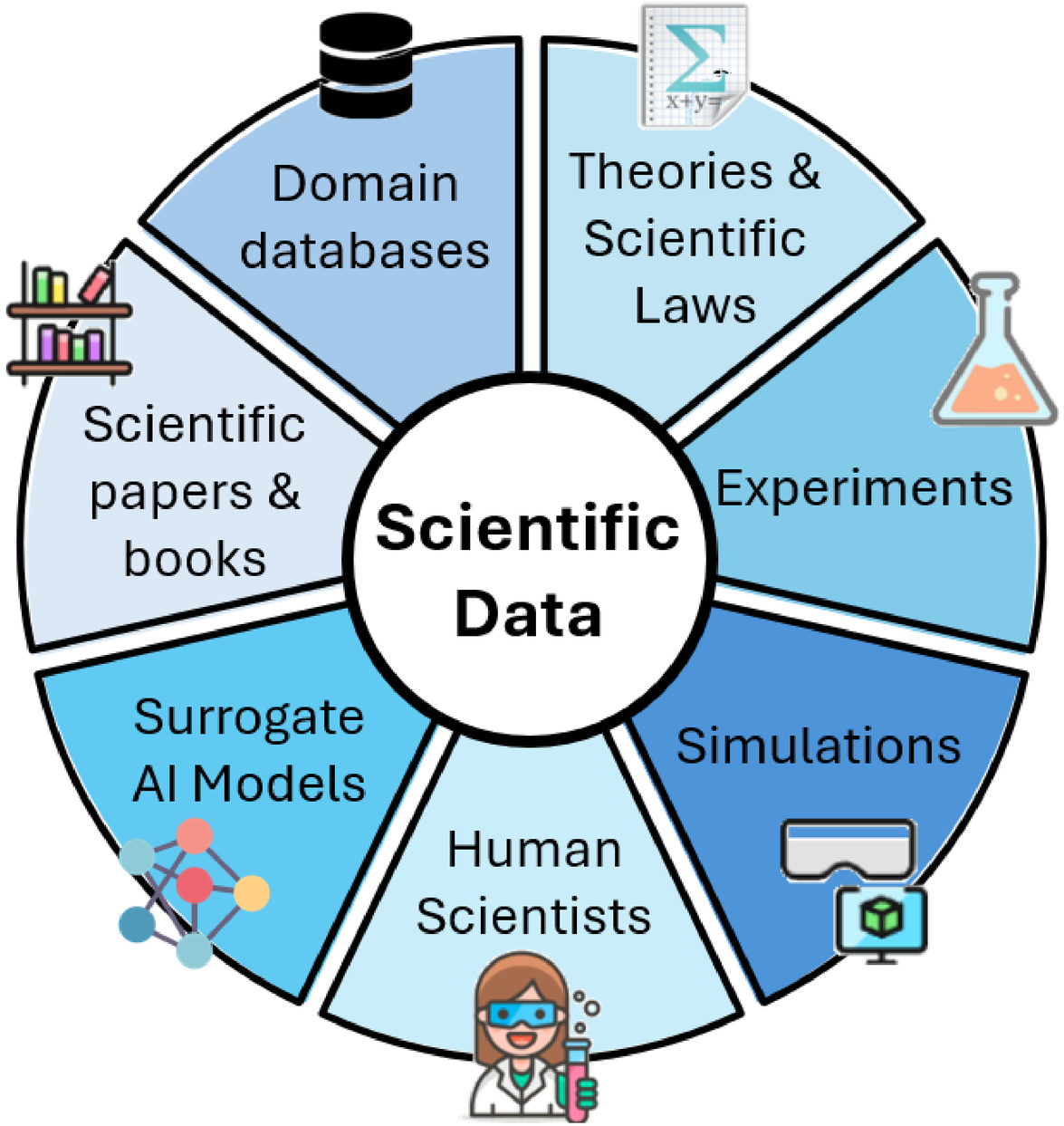
Reddy and Shojaee, 2025

LLM-guided Data-driven Discovery



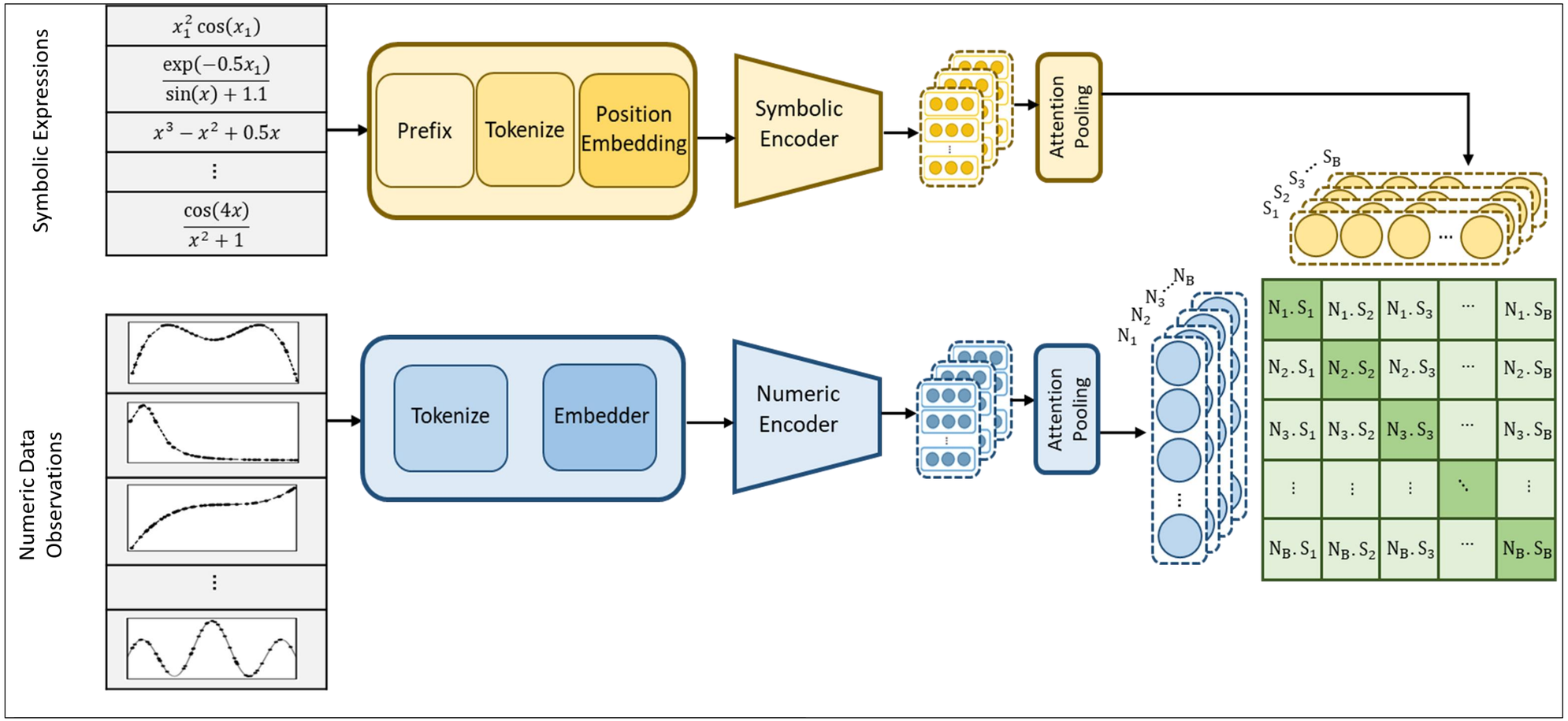
* “LLM-SR: Scientific Equation Discovery via Programming with Large Language Models” Shojaee et al., ICLR 2025 Oral

Diverse Nature of Scientific Data



Multimodal Representation Learning in Science

SNIP: Symbolic Numeric Integrated Pre-training in Mathematics



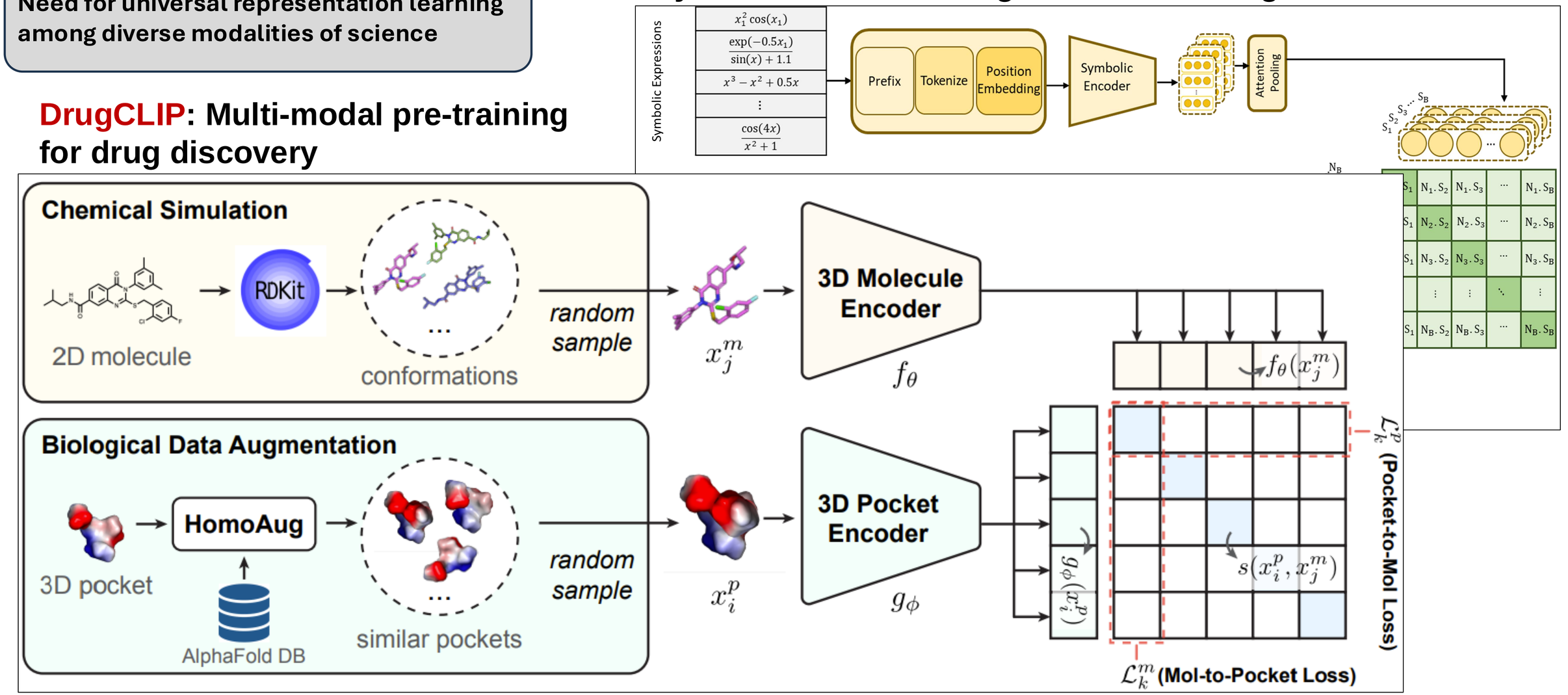
* SNP, Meidani et al., ICLR 2024 Spotlight

Multimodal Representation Learning in Science

Need for universal representation learning among diverse modalities of science

SNIP: Symbolic Numeric Integrated Pre-training in Mathematics

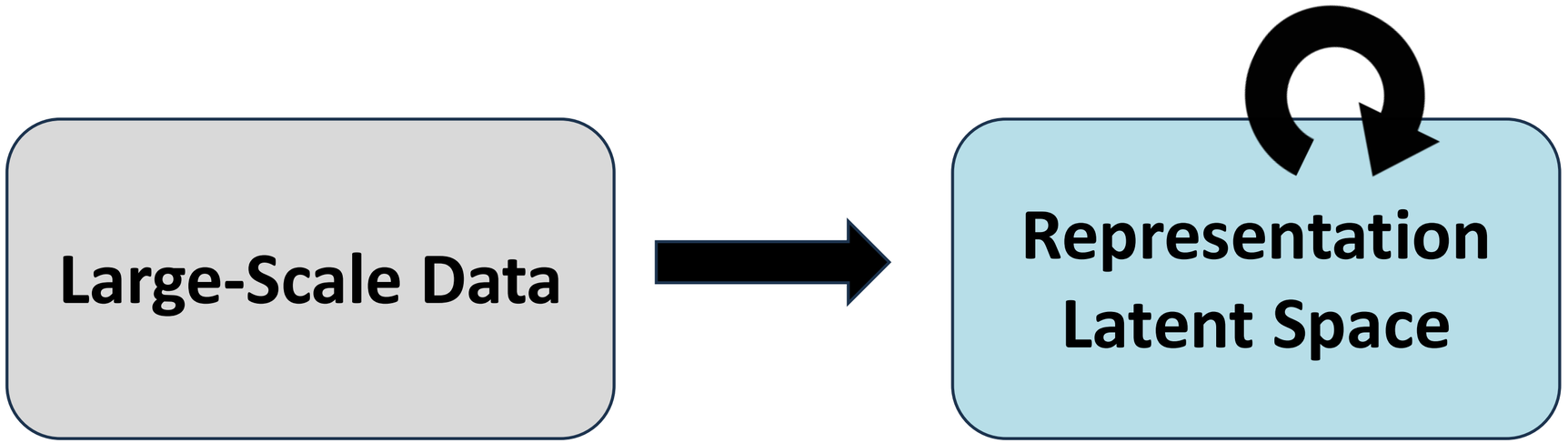
DrugCLIP: Multi-modal pre-training for drug discovery



* DrugCLIP, Gao et al., NeurIPS 2023

Latent Space Scientific Hypothesis Search

- Search for hypotheses in the combinatorial spaces ✘ Discrete / Combinatorial , ✘ Very Large, ✘ Inefficient
- **Move the search into learned latent spaces** ✓ Continuous, ✓ Low-dimensional, ✓ Interpolatable

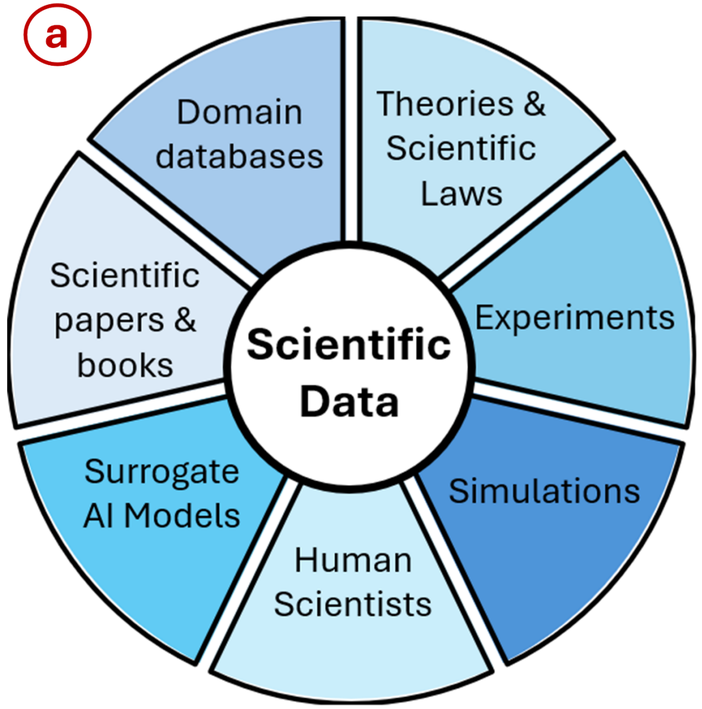


- Multi-modal
- Self-supervised
- Generative

…

Science-focused AI Agents

a



Inputs: b

- Scientific Task / Question
- Multi-modal Scientific Data
- Relevant Retrieved Scientific Context from Literature

Action: c

Tasks

- Hypothesis Generation
- Experiment Design
- Multi-modal Data Analysis
- Data Development

Tools

- API Calls
- Domain Simulators
- AI Surrogate Models
- Experiments

Evaluation: d

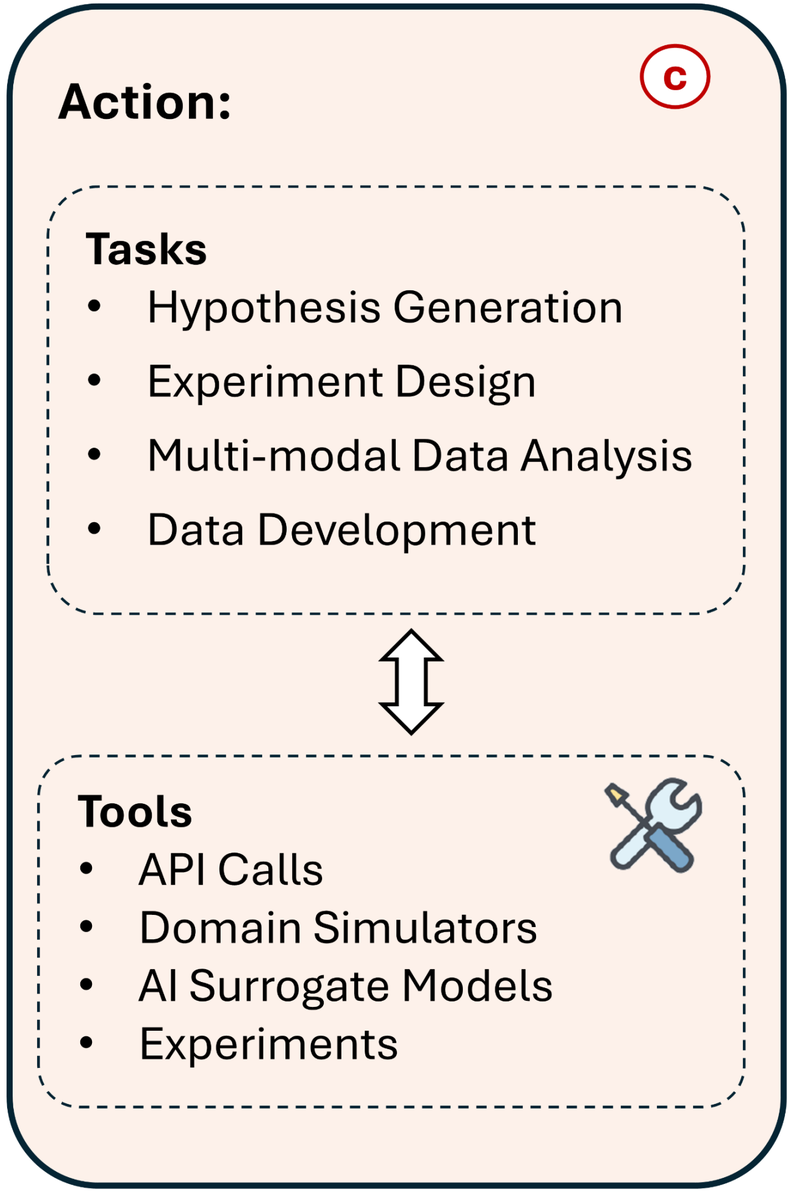
Metrics

- Data-driven Accuracy
- Generalizability
- Alignment with Scientific Principles
- Novelty

Tools

- Simulations
- Experiment Execution
- Human Scientists

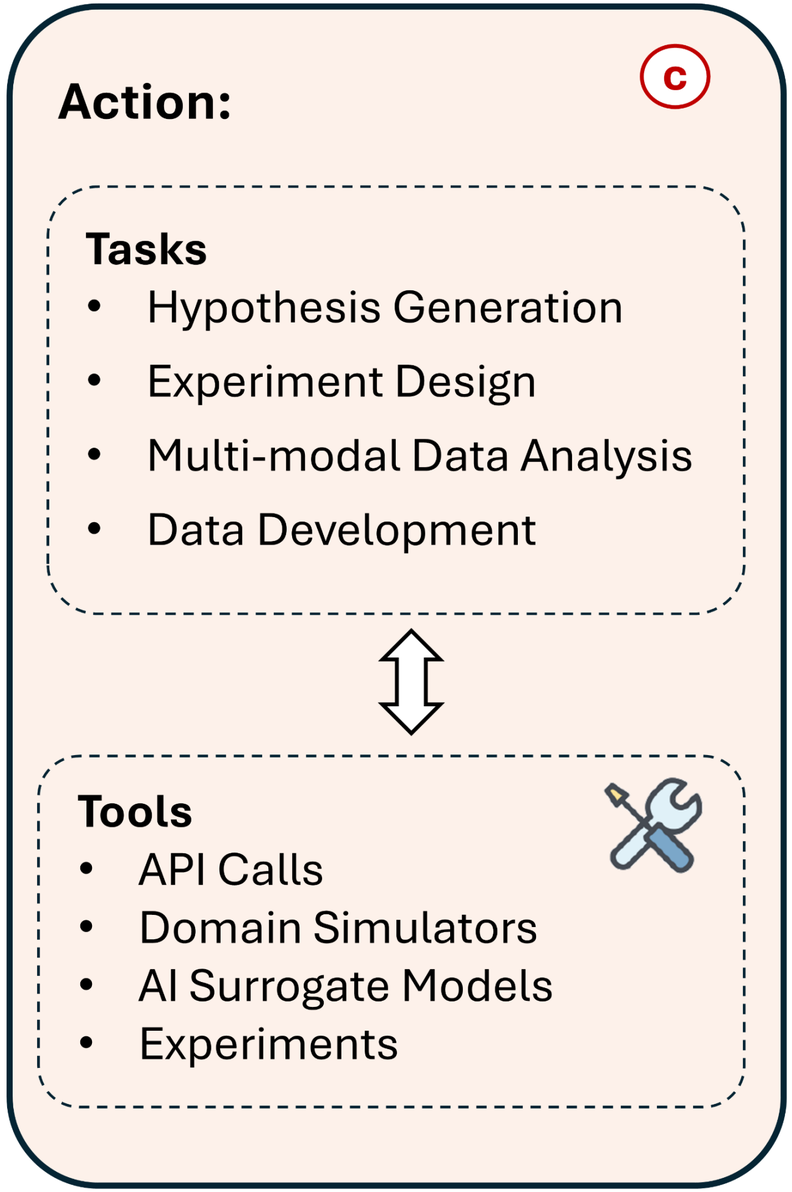
Science-focused AI Agents



◉ Actions:

- **Hypothesis Generation** → suggests potential scientific hypotheses by analyzing existing literature, patterns in data, and prior knowledge.
- **Experiment Design** → optimizes the design of scientific experiments by choosing parameters that maximize information gain.
- **Multi-modal Data Analysis** → understands data in other modalities (beyond text) with the help of programming execution
- **Data Generation** → generates synthetic data or simulations to explore scenarios before real-world experimentation.

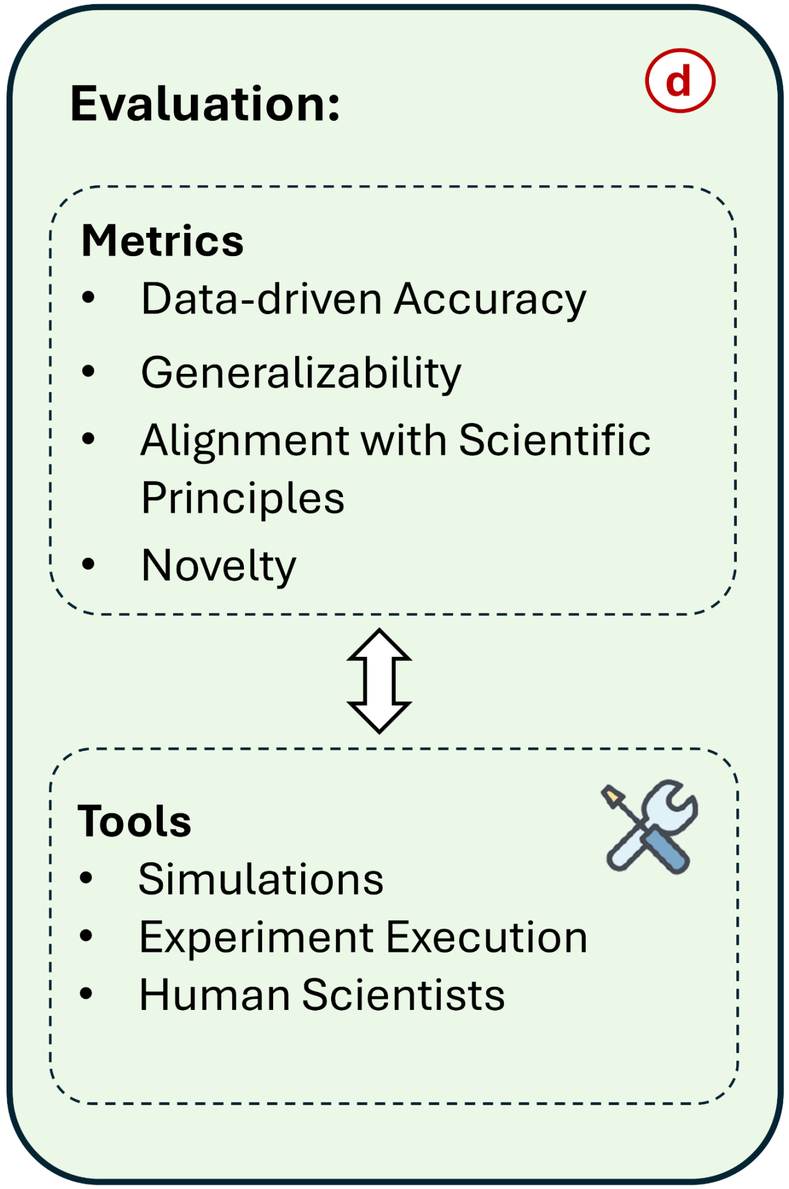
Science-focused AI Agents



◉ Tools:

- **API Calls** → Connects to databases (e.g., PubMed, NASA) for real-time knowledge.
- **Domain Simulators** → Runs physics, chemistry, and biology simulations.
- **AI Surrogate Models** → Replaces expensive simulations with fast, trained neural models.
- **Code Execution** → Runs Python, automates reasoning, and verifies results dynamically.

Science-focused AI Agents



◉ Metrics:

AI-driven discoveries must be **rigorously evaluated** to ensure they are reliable and impactful:

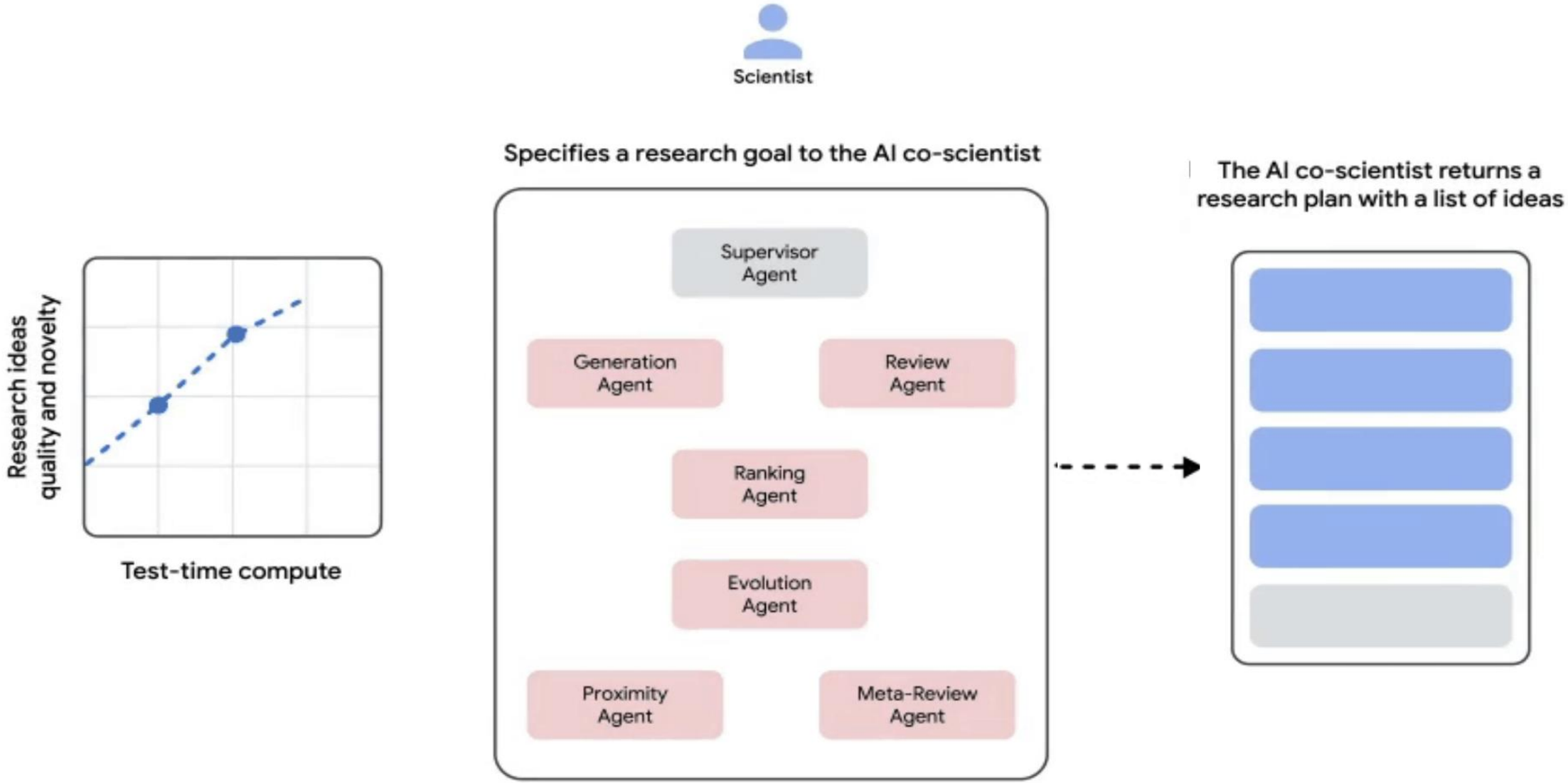
- **Data-driven Accuracy** → Does the AI's output match real-world experimental results?
- **Generalizability** → Can the discovery apply to new, unseen data?
- **Alignment with Scientific Principles** → Does it obey known physical laws and logical consistency?
- **Novelty** → Is it truly new, or just rediscovering existing knowledge?

Science-focused AI Agents

🚀 **Takeaway:**

LLM agents are **not just passive assistants**—they act as **programmable scientific collaborators**, integrating reasoning, execution, and data analysis beyond text to accelerate discovery.

Example: AI Co-Scientist



* Gottweis et al., 2025 (Google Deepmind)

Key Challenges & Research Opportunities

◉ Benchmarks for Scientific Discovery

Developing benchmark datasets focused on novel scientific discovery rather than recovery of known concepts

- **Current benchmarks are limited:**
  - Based on college-level scientific question-answering problems → does not capture the complexity of scientific discovery process
  - Vulnerable to reciting or memorization by LLMs → leading to overestimation of true discovery capabilities

Challenge? Balancing the use of LLMs' prior knowledge while avoiding mere recitation or memorization

Key Challenges & Research Opportunities

◉ Benchmarks for Scientific Discovery

Creating evaluation metrics for multiple facets of scientific discovery

- **Key Metrics:**
  - **Novelty:** quantify how different a discovered hypothesis or law is from existing knowledge.
  - **Generalizability:** assess how well discovered laws or models predict out-of-distribution unobserved data.
  - **Alignment with Scientific Principles:** evaluate whether discovered hypotheses are consistent with fundamental laws of physics or other well-established scientific principles.

Key Challenges & Research Opportunities

◉ Theory & Data Unification

In Natural Sciences

Scientific discovery typically involves **a complex interplay between theoretical derivation/reasoning with empirical data-driven observations**

In AI / ML

Scientific tasks focus on just one of these aspects (theorem proving & data-driven modeling in isolation) → **Mostly only on data-driven modeling**

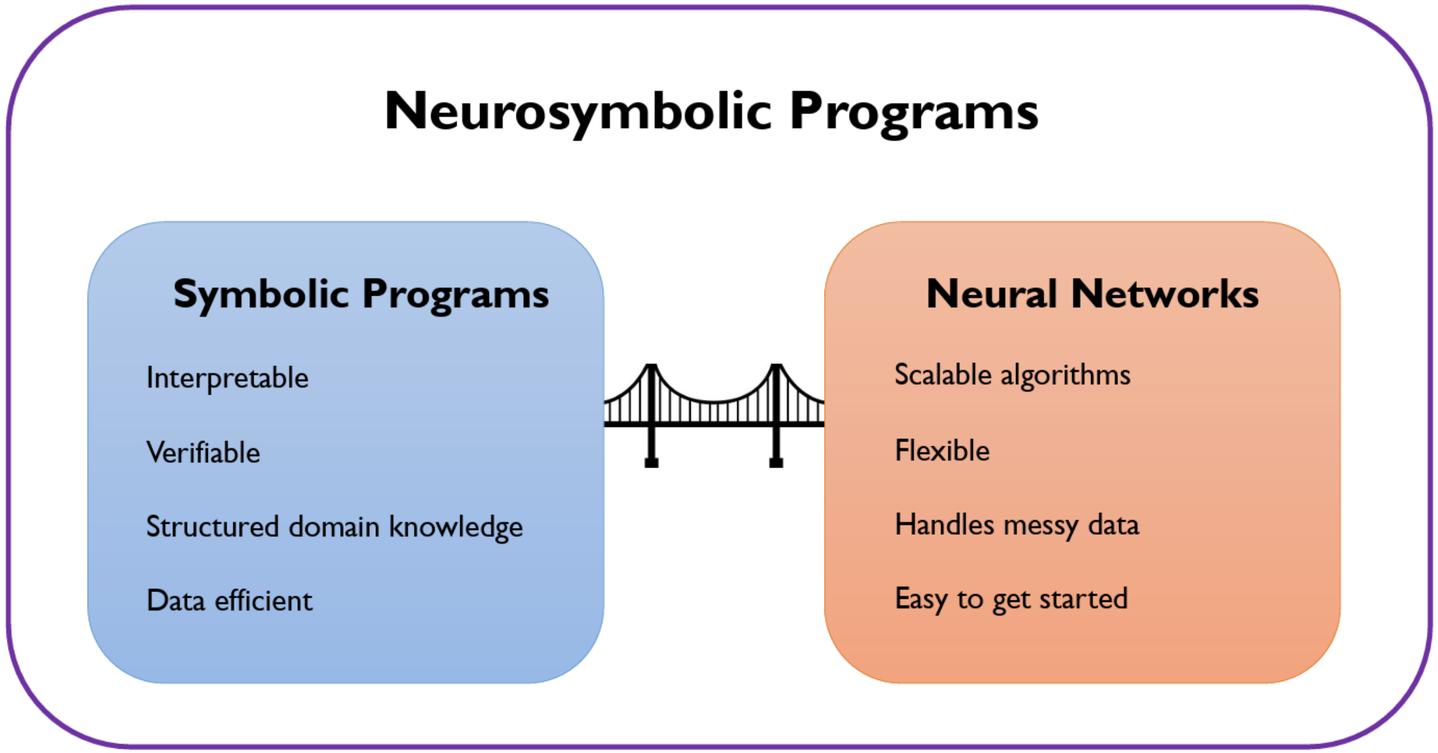
There is a pressing need for unified frameworks that **integrate logical and mathematical reasoning, formal theorem proving, and data-driven modeling**

Key Challenges & Research Opportunities

◉ Neuro-Symbolic AI

How to effectively integrate the strengths of symbolic reasoning (e.g., logical deduction, formal proofs) with the flexibility and learning capabilities of neural networks?

Transition between symbolic and neural representations is helpful in capturing full spectrum of scientific reasoning!!



* Plot borrowed from “Neuro-Symbolic Programming” tutorial, NeurIPS 2022

Conclusion

◉ Now is a **pivotal moment** in AI for Scientific Discovery!!

◉ Recent advancements in LLMs, multimodal learning, and AI-driven reasoning have demonstrated great promise in accelerating scientific progress.

◉ **Key Challenges and Research Opportunities Ahead**

- **Developing benchmarks** to evaluate true scientific discovery, beyond memorization.
- **Theory & Data Unification** to integrate formal and logical reasoning, theoretical derivation along with data-driven modeling in the discovery process.
- **Advancing science-focused AI agents** that autonomously design, verify, and refine scientific hypotheses.
- **Enhancing multimodal representation learning in science** to incorporate scientific data from diverse modalities effectively & efficiently into the discovery process.

Our Team

The 39th Annual AAAI Conference on Artificial Intelligence

FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, PENNSYLVANIA, USA

Chandan Reddy

Parshin Shojaee

Link to Paper:

Collaborators:

Kazem Meidani

Amir Barati Farimani

Shashank Gupta

Nour Makke

Sanjay Chawla

VIRGINIA TECH

CARNEGIE MELLON UNIVERSITY PITTSBURGH PENNSYLVANIA 1900

CMU

AI2

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Thank you!

Website: http://creddy.net

Github Repo: https://github.com/deep-symbolic-mathematics/