5

Science

Overview

The speed with which AI is transforming science is accelerating, with momentum from the 2024 Nobel Prize in Chemistry awarded to Dennis Hassabis, John Jumper, and David Baker for their work on AI-driven protein structure prediction and design, and the operational deployment of AI weather models at the European Centre for Medium-Range Weather. In 2025, AI moved beyond improving individual pipeline steps and toward replacing entire scientific workflows, from weather prediction to multiagent hypothesis generation and experimental design. Still, rigorous benchmarks continue to expose large gaps between plausible output and reliable scientific work, with frontier agents scoring below 20% on paper-scale replication tasks. AI's impact in social sciences has been slower to emerge but with notable exceptions. Linguistics and computational language research helped lay the groundwork for modern language models, and those models are now being applied back into fields such as linguistics, communications, and network analysis. Progress in these areas is harder to capture through the datasets and benchmarks covered in this chapter.

Contents

Chapter Highlights

1. **AI-related scientific publications are growing year-over–year.** Natural sciences reached approximately 80,150 AI publications in 2025, up 26% from 2024. AI now accounts for 5.8%–8.8% of scientific research output depending on the field, up from below 1% in 2010.

2. **Frontier models outperform human chemists on average but cannot reproduce published research.** On ChemBench, the best models surpass human expert averages across 2,700+ chemistry questions while struggling with basic tasks. On ReplicationBench, frontier models score below 20% on paper-scale replication in astrophysics. On UnivEarth, LLM agents answer Earth observation questions with 33% accuracy, and their code fails 58% of the time.

3. **Astronomy released its first foundation model, first visualization benchmark, and a 100TB training dataset in 2025, signaling a field-wide shift toward AI infrastructure.** AION-1, trained on over 200 million celestial objects from 5 major surveys, is the first astronomy foundation model. AstroVisBench introduced the first benchmark for LLM scientific computing and visualization in the field.

4. **An AI system ran a full weather forecasting pipeline end-to-end for the first time in 2025.** Aardvark Weather replaced the traditional numerical prediction pipeline with a single ML system, and multiple AI weather models reached operational deployment. FourCastNet 3 generates a 60-day global forecast in under 4 minutes, running 8 to 60 times faster than prior approaches.

5. **On end-to-end scientific research tasks, the best AI agents score roughly half of what PhD experts achieve.** On PaperArena, the best agent reaches 38.8% accuracy versus a PhD expert baseline of 83.5%. On BixBench, frontier models achieve roughly 17% accuracy on real-world bioinformatics analysis.

6. **The first fully AI-generated paper was accepted at a peer-reviewed workshop in 2025, but the list of experimentally confirmed AI discoveries remains short.** Sakana's AI Scientist-v2 produced a paper accepted at an ICLR workshop without human-coded templates. Google's AI Co-Scientist was validated in three biomedical areas.

7. **Most AI models for science originate from academic and government institutions, in contrast with the industry-dominated landscape of general-purpose AI.** Many AI foundation models for science result from international collaborations. Earth science datasets come entirely from government and academic sources, while industry leads foundation model development in weather and climate.

5.1 AI for Science in 2025

AI's role in science falls into three categories that coexist but differ in terms of maturity. The first—machine learning over scientific data to build predictive and explainable models—has been practiced for several decades and is now commonplace. The second—AI systems that assist scientists in their workflows through literature synthesis, experiment design, or data analysis—has been emerging over several years and expanded considerably in 2025. The third category—autonomous AI systems capable of generating new scientific discoveries with limited human guidance—is gaining traction but it remains at an early stage. The year's most visible developments occurred primarily in the second and third categories. Aardvark Weather replaced the full numerical weather prediction pipeline (Allen et al., 2025). Google's AI Co-scientist orchestrated hypothesis generation through experimental design (Gottweis et al., 2025). To date, the clearest breakthroughs tend to cluster in domains with strong existing data infrastructure, including structural biology, physics, chemistry, and materials science, rather than in fields with the most sophisticated mathematical or physics-based models.

These developments, however, do not automatically translate into scientific progress. Experimental validation remains expensive and time-consuming, and scientists are unlikely to invest in testing AI-generated hypotheses without sufficient reason to believe it will yield some findings. In drug discovery, for example, AI systems can propose novel candidate molecules at scale, but clinical trials to determine whether those molecules work remain a costly, multiyear process. The gap between what AI can propose and what scientists can feasibly test is a recurring theme across the domains covered in this chapter.

Publications in AI for Science

In the Web of Science database, AI-related publications in the natural sciences reached approximately 80,150 in 2025, up from 63,547 in 2024, a one-year increase of roughly 26% (Figure 5.1.1). Physical sciences[1] and life sciences followed similar trajectories in 2025, reaching approximately 33,000 and 29,000 publications, respectively, with each growing by roughly 27%–28% year over year. Earth science—the smallest category in absolute terms at approximately 20,460 publications—grew by about 23%. As a share of total scientific output, AI-related work remains a single-digit fraction of each field but is climbing quickly (Figure 5.1.2). By 2025, Earth science had the highest AI penetration at 8.8%, followed by natural sciences overall at 6.8%, life sciences at 6.5%, and physical sciences at 5.8%. In 2010, all four categories sat below 1%. The quantity of AI-mentioning papers is not the same as the quality of AI-enabled discovery, but the breadth suggests that AI methods are becoming a routine part of scientific practice across disciplines.

1 Physical sciences in this analysis include: astronomy and astrophysics, chemistry, crystallography, electrochemistry, geochemistry and geophysics, geology, mathematics, meteorology and atmospheric science, mineralogy, mining and mineral processing, oceanography, optics, physical geography, physics, polymer science, thermodynamics, and water resources.

Number of AI-related publications in natural sciences, 2010–25

Source: AI Index, 2026 | Chart: 2026 AI Index report

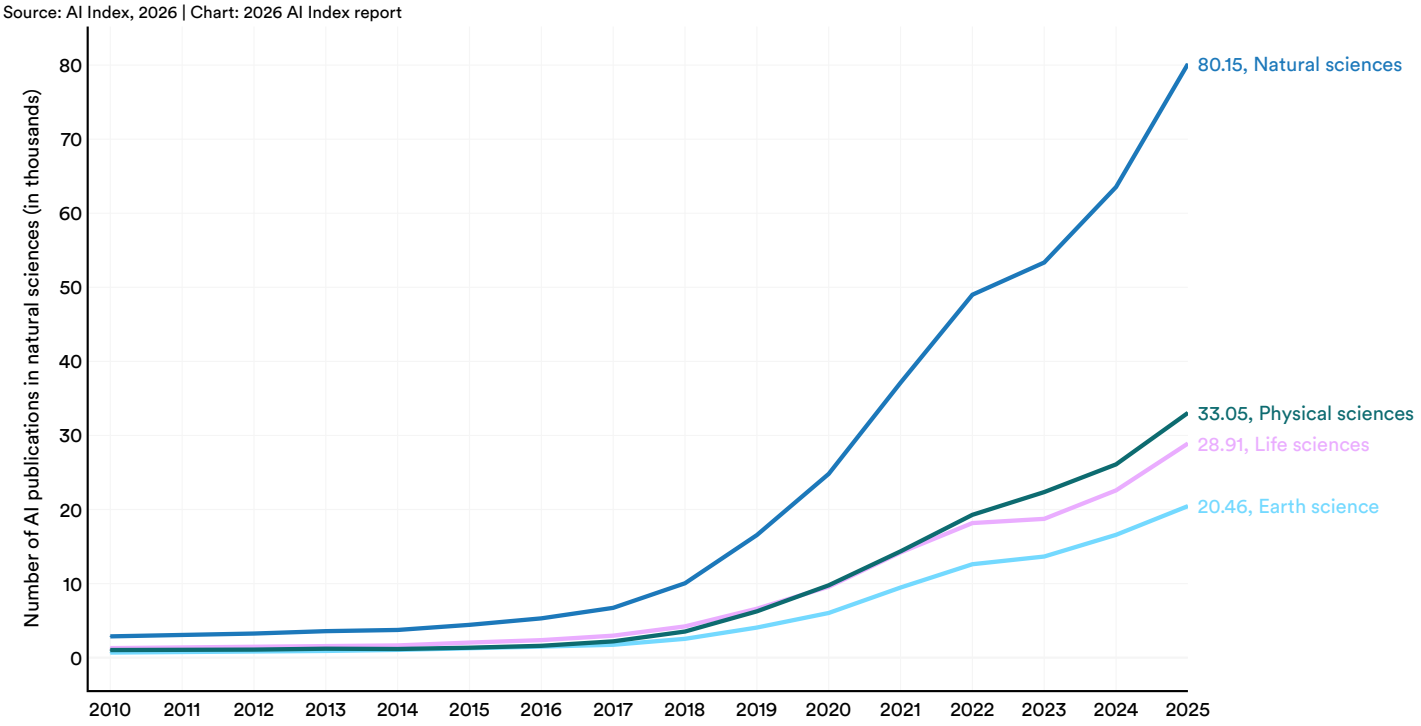


Figure 5.1.1[2]

AI-related publications in natural sciences (% of total), 2010–25

Source: AI Index, 2026 | Chart: 2026 AI Index report

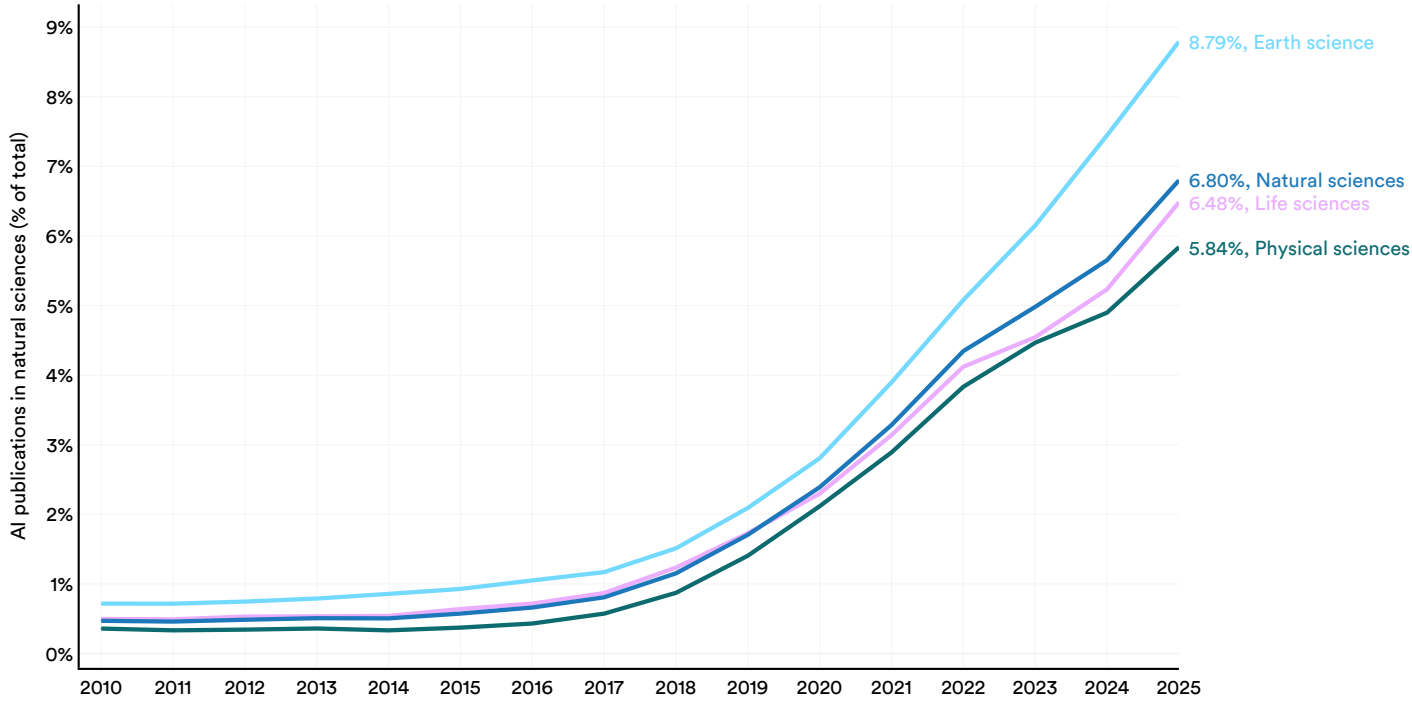


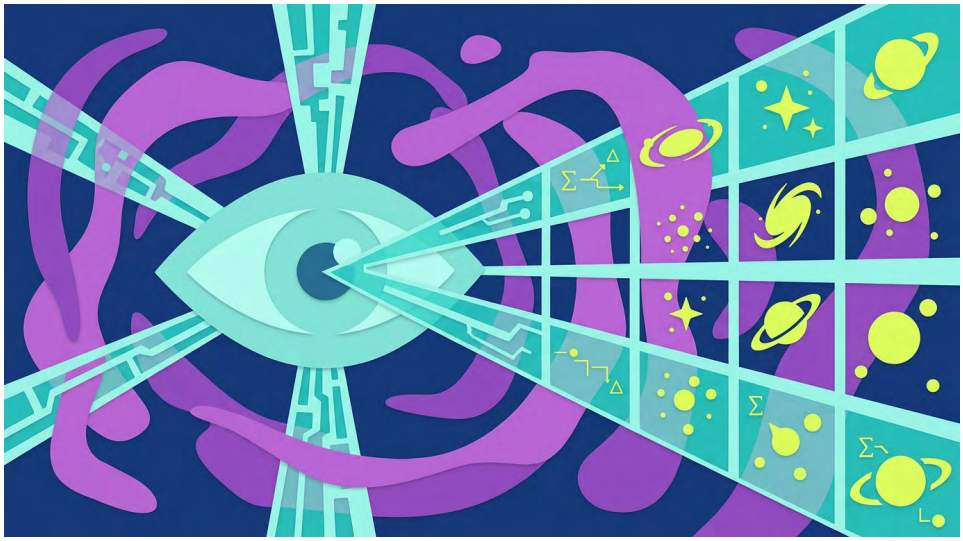
Figure 5.1.2

2 The natural sciences count may be slightly lower than the sum of the individual domain counts. This is because a single publication can be assigned to more than one domain. For example, a biochemistry paper may be categorized under both physical sciences and life sciences. To avoid double-counting, these publications are counted only once in the natural sciences total.

5.2 AI Across Scientific Domains

This section examines AI's expanding role in science across three major scientific groupings and tracks the datasets, benchmarks, and foundation models of each. The tables below catalog selected releases across each category. A consistent finding is that the majority of scientific AI models originate from academic institutions collaborating across countries, in contrast to the industry-dominated landscape of general-purpose foundation models described in Chapter 1 and Chapter 2.

Physics, Astronomy, Chemistry, and Materials Science



AI is accelerating physics, astronomy, chemistry, and materials science by replacing expensive first-principles simulations with learned surrogates and by generating novel materials and molecular structures through inverse design. Notable 2025 releases include large chemistry datasets (e.g., OMol25 and OC25), simulation-oriented foundation models (e.g., Walrus, GPhyT), and materials checkpoints for atomistic modeling and generation (e.g., MACE-MP-0 and MatterGen). In chemistry and materials science, agent systems began connecting to external software tools and laboratory equipment to execute experiments. Benchmark results, however, suggest these systems are not yet reliable when asked to carry out full research tasks from start to finish.

Datasets

The largest dataset releases in physics and chemistry in 2025 expanded multimodal astronomy resources and chemistry-scale quantum data. Multimodal Universe aggregates approximately 100TB of astronomical observations, while OMol25 reports over 100 million high-accuracy density functional theory (DFT) calculations spanning 83 elements. These datasets provide the training foundation for large models targeting prediction and simulation tasks in their respective fields (Figure 5.2.1).

Selected datasets in physics, astronomy, chemistry, and materials science (2025)

Name	Affiliation[3]	Domain	Sector	Summary
ChemPile	Helmholtz Institute for Polymers (HIPOLE) Jena, Friedrich Schiller University Jena, Hacettepe University, University of Toronto	Chemistry	ACADEMIA, NONPROFIT	75B-plus tokens of curated chemical data (SMILES, InChI, text, code, educational materials).
Multimodal Universe	Polymathic AI, Instituto de Astrofisica de Canarias, Universidad de La Laguna, Massachusetts Institute of Technology	Astronomy	ACADEMIA, NONPROFIT	100TB astronomical dataset: multichannel images, spectra, time series from hundreds of millions of observations.
The Open Molecules 2025 (OMol25)	Fundamental AI Research (FAIR) at Meta, Los Alamos National Laboratory, University College Dublin	Chemistry, Materials, Chemical Physics	INDUSTRY, GOVERNMENT, ACADEMIA	100M-plus DFT calculations spanning 83 elements, diverse chemical interactions, structures up to 350 atoms.
The Open Catalyst 2025 (OC25)	FAIR at Meta, Texas Tech University, Nanyang Technological University	Chemistry, Materials	INDUSTRY, GOVERNMENT, ACADEMIA	7.8M calculations across 1.5M solvent environments spanning 88 elements for catalysis at solid-liquid interfaces.

Figure 5.2.1

Benchmarks

In these particular domains, benchmarks have been newly introduced and therefore do not offer longitudinal data across multiple years. It is interesting to see how general-purpose frontier models, discussed in Chapter 2, perform on scientific tasks (Figure 5.2.2). On ChemBench, a chemistry evaluation with over 2,700 question-answer pairs, the best frontier models outperform the best human chemists, though they struggle with basic tasks. ReplicationBench reports frontier model performance below 20% on paper-scale replication tasks in computational astrophysics.

3 Full references are provided in the Appendix.

Selected benchmarks in physics, astronomy, chemistry, and materials science (2025)

Name	Affiliation[4]	Domain	Sector	Summary
AstroVisBench	University of Texas at Austin, NSF National Optical-Infrared Astronomy Research Laboratory, University of Virginia	Astronomy	ACADEMIA, GOVERNMENT	First benchmark for LLM scientific computing and visualization in astronomy.
Chembench	Friedrich Schiller University Jena, Helmholtz Institute for Polymers (HIPOLE) Jena, Spanish National Research Council (CSIC)	Chemistry	ACADEMIA, INDUSTRY, GOVERNMENT	2,700-plus Q&A pairs. Best models outperform human chemists on average but struggle with basic tasks.
ChemX	ITMO University, D ONE	Chemistry, Materials Science	ACADEMIA, INDUSTRY	10 curated datasets for automated chemical information extraction from nanomaterials and small molecules.
GravityBench	University of Toronto, NYU Abu Dhabi	Physics, Astrophysics	ACADEMIA	Tests AI discovery of physics laws from gravitational simulations, including non-real-world physics.
LLM-SRBench	Virginia Tech, VinUniversity, Carnegie Mellon University	Physics, Scientific Equation Discovery	ACADEMIA, INDUSTRY	239 problems testing genuine equation discovery vs. memorization. Best systems score 31.5%.
Matbench Discovery	University of Cambridge, Lawrence Berkeley National Laboratory, Federal Institute of Materials Research and Testing (BAM)	Materials, Chemistry	ACADEMIA, GOVERNMENT	Evaluation framework for machine learning energy models prescreening stable inorganic crystals.
MatSciBench	UCLA, Princeton, Virginia Tech	Materials Science	ACADEMIA	1,340 college-level problems across 6 fields and 31 subfields. Top models under 80%.
PHYBench	Peking University, CSRC	Physics	ACADEMIA	500 original physics problems. Gemini 2.5 Pro: 36.9% vs. human experts: 61.9%.

4 Full references are provided in the Appendix.

PhysGym	KAUST Center of Excellence for Generative AI, The Swiss AI Lab	Physics	ACADEMIA, INDUSTRY	Interactive physics environments testing LLM scientific reasoning under varying prior knowledge.
ReplicationBench	Stanford University, University of Toronto	Astrophysics/ Research Replication	ACADEMIA	Tests AI replication of entire astrophysics papers. Frontier models score under 20%.
TheoreticalPhysics Benchmark (TPBench)	University of Wisconsin-Madison, Indiana University, NSF-Simons AI Institute for the Sky (SkAI)	Theoretical Physics	ACADEMIA	57 novel theoretical physics problems (high-energy theory, cosmology). Research-level problems largely unsolved.

Figure 5.2.2

Foundation Models

Foundation model (FM) releases in 2025 spanned astronomy, physics simulation, chemistry language models, and materials modeling (Figure 5.2.3). GPhyT, a General Physics Transformer, trained on 1.8TB of simulation data, achieved up to 29 times better performance than specialized models, and generalized to physics problems outside its training data without task-specific fine-tuning.

Selected foundation models in physics, astronomy, chemistry, and materials science (2025)

Name	Affiliation[5]	Domain	Sector	Summary
AION-1	UC Berkeley, Flatiron Institute, New York University	Astronomy	ACADEMIA, NONPROFIT, GOVERNMENT	Astronomy FM: 300M–3.1B parameters, 200M-plus celestial objects from 5 major surveys. Open release.
ChemDFM	Shanghai Jiao Tong University, Suzhou Laboratory, AI Speech Co.	Chemistry	ACADEMIA, GOVERNMENT, INDUSTRY	Chemistry LLM: 34B tokens, 2.7M instructions. Generalist chemical AI.

5 Full references are provided in the Appendix.

GPhyT: General Physics Transformer	RWTH Aachen University, University of Virginia	Physics	ACADEMIA	Trained on 1.8 TB simulation data. Up to 29x better than specialized models. Zero-shot generalization.
MACE-MP-0	University of Cambridge, Federal Institute of Materials Research and Testing (BAM), UC Berkeley	Chemical Physics	ACADEMIA, GOVERNMENT, INDUSTRY	General-purpose force field model for predicting atomic interactions across nearly all materials.
MatterGen	Microsoft Research AI for Science, Shenzhen Institute of Advanced Technology (Chinese Academy of Sciences)	Materials Science	INDUSTRY, GOVERNMENT	Diffusion-based generative model. Over 2x more novel and stable than existing methods.
PDE-Transformer	Technical University of Munich	Physics Simulations	ACADEMIA	Transformer for physics partial differential equations (PDEs) on grids. Outperforms state-of-the-art vision architectures across 16 types of physics simulations.
PhysiX	UCLA	Physics Simulations	ACADEMIA	4.5B params. First large-scale physics simulation FM. Transfers from natural videos to simulation.
SMI-TED	IBM Research	Chemistry	INDUSTRY	Chemical foundation models trained on molecular sequences.
Surya	University of Alabama in Huntsville, NASA Marshall Space Flight Center, IBM Research	Heliophysics	ACADEMIA, GOVERNMENT, INDUSTRY	366M parameters. First heliophysics FM. Forecasts space weather from NASA's Solar Dynamics Observatory data without task-specific training.
Walrus	Flatiron Institute, New York University, University of Cambridge	Fluid Mechanics Continuum Mechanics, multiple domains	ACADEMIA, NONPROFIT, INDUSTRY	Fluid mechanics FM: 19 scenarios spanning astrophysics, geoscience, plasma physics, acoustics. Open weights.

Figure 5.2.3

AI Agents

Agent systems in the physical sciences combine tool use with domain-specific reasoning to perform tasks requiring multiple steps. Some of these systems function as focused components within larger pipelines, while others attempt to operate as end-to-end research systems. As they take on more responsibility for both designing and executing research, independent confirmation of results becomes an important step.

Physics Supernova scored 23.5 out of 30 at the 2025 International Physics Olympiad, ranking 14th out of 406 participants and reaching gold-medalist level. StarWhisper Telescope automates astronomical observation planning across 10 telescopes. In chemistry, ChemAgents demonstrated autonomous synthesis and optimization using a robotic platform controlled by Llama-3.1-70B (Figure 5.2.4).

Selected AI agents in physics, astronomy, chemistry, and materials science (2025)

Name	Affiliation[6]	Domain	Sector	Summary
ChatGPTMaterial Explorer	John Hopkins University	Materials Science	ACADEMIA	Materials science assistant combining LLMs with graph neural networks for property prediction.
ChemAgents	University of Science and Technology of China, University of Birmingham, Henan Academy of Sciences	Chemistry	ACADEMIA, GOVERNMENT	Robotic AI chemist (Llama-3.1-70B). Autonomous synthesis, optimization, photocatalysis.
ChemToolAgent	Shanghai Artificial Intelligence Laboratory, Soochow University, Zhejiang University	Chemistry, Materials Science	ACADEMIA, GOVERNMENT	137 external chemical tools. HE-MCTS framework surpasses GPT-4o on chemistry QA.
Crystalyse	Imperial College London	Materials Science, Chemistry	ACADEMIA	Multi-tool AI agent for materials design that coordinates multiple computational tools through an LLM-based reasoning framework.
Physics Supernova	Princeton University, Tsinghua University, Shanghai Jiao Tong University	Physics	ACADEMIA	IPhO 2025: 23.5 out of 30, ranked 14th of 406. Gold-medalist-level physics problem-solving.
StarWhisper Telescope	University of Chinese Academy of Sciences, National Astronomical Observatories (CAS), Simon Fraser University	Astronomy	ACADEMIA, GOVERNMENT, INDUSTRY	Automates astronomical observations across 10 telescopes. LLM-driven observation planning.

Figure 5.2.4

6 Full references are provided in the Appendix.

Biological and Life Sciences

AI is increasingly being applied to biological research beyond biomedicine to address fundamental questions in genomics, neuroscience, ecology, and synthetic biology. Chapter 6 covers AI's role in the therapeutic pipeline, from protein structure prediction to drug design to clinical applications. This section focuses on the broader scientific infrastructure, including the datasets, benchmarks, foundation models, and agents that support biological research as a whole.

The scale of biological training data grew in 2025, and foundation models trained on genomic and evolutionary data expanded from prediction into generative design. However, the gap between genomic sequence data, which is abundant, and functional perturbation data, which measures how biological systems respond to interventions, remains wide (Callahan et al., 2025; Sun et al., 2025). AI is also being applied at the macroscopic scale, with computer vision and acoustic models routinely processing sensor data to track species populations and optimize agricultural water use in real time (Miller et al., 2025; Khan and Sharma, 2025). Ecological and biodiversity applications lag behind other biological subfields, in part because training data in these areas is sparse, biased toward well-studied taxa, and lacking standardized formats (Fahsbender et al., 2025). In species taxonomy and evolutionary biology, vision-based foundation models such as the BioCLIP family (Stevens et al., 2024; Gu et al., 2025) are enabling classification and discovery across the tree of life, while methods like PhyloNN (Elhamod et al., 2023) can identify evolutionary traits from images without labeled data.

In neuroscience, AI serves both as a practical tool for brain mapping and as a source of theoretical inspiration. Computer vision approaches have been instrumental in assembling full connectome data from model organisms such as the fly and mouse (Dorkenwald et al., 2025; The MICrONS Consortium, 2025). Comparisons between biological neuronal networks and artificial deep networks inform how researchers understand the principles of information processing in the brain (Linsley et al., 2025; Kazemian et al., 2025).

Datasets

OpenGenome2 contains nearly 9.3 trillion base pairs of curated DNA from across all domains of life, making it the largest genomic training corpus assembled to date and the foundation for the Evo 2 model. In neuroscience, Spacetop contributed over 600 imaging hours across 101 participants for cognitive neuroscience research. These resources provide the scale necessary for foundation models to learn biological features without task-specific training, though the gap between genomic data availability and functional perturbation data remains wide (Figure 5.2.5).

Selected datasets in biological and life sciences (2025)

Name	Affiliation[7]	Domain	Sector	Summary
OpenGenome2	Arc Institute, Stanford University, Nvidia	Biology, Genomics	ACADEMIA, NONPROFIT, INDUSTRY	9.3T base pairs of curated DNA from all domains of life. Training corpus for Evo 2.
ProteinTalks-DB	Westlake University (Academia)	Proteomics, Systems Biology		
Spacetop	Dartmouth College, Johns Hopkins University, Emory University	Neuroscience	ACADEMIA	101 participants, 600-plus imaging hours. Cognitive, affective, social, interoceptive domains.

Figure 5.2.5

Benchmarks

Life-science benchmarks have moved to testing workflow execution and tool-integrated analysis, rather than static knowledge. BixBench reports that frontier models achieve roughly 17% accuracy on real-world bioinformatics analysis tasks, highlighting challenges in chaining tools, file handling and domain interpretation. BioML-bench provides the first end-to-end evaluation of AI agents on biomedical machine learning tasks that span protein engineering to drug discovery, and it found that on average agents underperform human baselines. These results are consistent with the pattern observed across other scientific domains in this chapter. AI systems perform well on isolated subtasks but struggle when required to execute the multistep workflows that actual biological research demands (Figure 5.2.6).

7 Full references are provided in the Appendix.

Selected benchmarks in biological and life sciences (2025)

Name	Affiliation[8]	Domain	Sector	Summary
BaisBench	Tsinghua University	Biology	ACADEMIA	Evaluates AI biological discovery ability via cell annotation and data-driven questions.
BioML-bench	Shift Bioscience, University of Cambridge, ScienceMachine	Biology	INDUSTRY, ACADEMIA	First end-to-end biomedical machine learning evaluation. Agents underperform human baselines on average.
BixBench	FutureHouse, ScienceMachine	Computational Biology	INDUSTRY	50-plus bioinformatics scenarios. GPT-4o and Claude 3.5 Sonnet: ~17% accuracy.
CGBench	Stanford University	Biology (Genetics)	ACADEMIA	Clinical genetics interpretation. Reasoning models excel at fine-grained tasks; substantial hallucination gaps remain.
Mouse vs. AI: Robust Foraging Competition	UC Santa Barbara	Neuroscience	ACADEMIA	Bioinspired benchmark grounding reinforcement learning agents in neuroscience via shared foraging tasks with mice.

Figure 5.2.6

Foundation Models

In 2025, foundation model releases in the biological and life sciences domains expanded across genomics and cellular modeling. Genomic foundation model Evo 2, which trained on OpenGenome2, trained on 9.3 trillion DNA base pairs from all domains of life. It operates at up to 40 billion parameters with a 1 million token context window and was released with fully open weights. Chapter 6 examines its performance on genomic prediction tasks alongside smaller, task-specific alternatives, and covers additional genomic and cellular foundation models, including AlphaGenome and CellFM. In neuroscience, a foundation model of neural activity predicts neuronal responses and generalizes across stimulus types and individual animals (Figure 5.2.7).

8 Full references are provided in the Appendix.

Selected foundation models in biological and life sciences (2025)

Name	Affiliation[9]	Domain	Sector	Summary
AlphaGenome	Google DeepMind	Biology, Genomics	INDUSTRY	Genomic foundation model predicting thousands of functional measurements from DNA sequence at single-base-pair resolution.
ANN Model	Baylor College of Medicine, Stanford University, University of Göttingen	Neuroscience	ACADEMIA	Neural activity FM. Predicts neuronal responses, generalizes across stimulus types and mice.
BioCLIP 2	The Ohio State University, Smithsonian Institution, UNC-Chapel Hill	Biology	ACADEMIA, NONPROFIT	Vision foundation model for biological classification across the tree of life. Trained using hierarchical contrastive learning on taxonomic structure.
BioLab	Princeton University, BioMap Research, Zhejiang University	Biology	ACADEMIA, INDUSTRY	Multiagent system for automated biological research. Experimentally validated novel antibody designs.
CellFM	Sun Yat-sen University, Chongqing University, Jinfeng Laboratory	Biology	ACADEMIA, GOVERNMENT, INDUSTRY	800M parameters, 100M human cells. Single-cell analysis, perturbation prediction, gene–gene relationships.
Evo 2	Arc Institute, Nvidia, Stanford University, UC Berkeley	Biology, Genomics	ACADEMIA, INDUSTRY	40B params, 1M token context. 9.3T base pairs. Genome-scale generation. Fully open release.
ProteinTalks	Westlake University, DP Technology Co., Westlake Omics Co.	Proteomics, Systems Biology	ACADEMIA, INDUSTRY	Foundation model for protein network dynamics. Predicts drug efficacy and synergy from perturbation proteome data.

Figure 5.2.7

9 Full references are provided in the Appendix.

AI Agents

Agent systems in the life sciences are beginning to operationalize complex research workflows, including literature synthesis and bioinformatics execution.

BCI-Agent performs autonomous neuronal cell-type classification from electrophysiology recordings without task-specific training. Biomni is a general-purpose biomedical agent spanning 25 subfields. Chapter 6 describes its architecture and capabilities in greater detail (See Figure 5.2.8).

Selected AI agents in biological and life sciences (2025)

Name	Affiliation[10]	Domain	Sector	Summary
BCI-Agent	Harvard University, Massachusetts Institute of Technology (MIT), Broad Institute of MIT and Harvard	Neuroscience	ACADEMIA	Autonomous neuronal cell-type classification from electrophysiology. No task-specific training.
BioAgents	UC Berkeley, Microsoft Research, UC San Francisco	Biology	ACADEMIA, INDUSTRY	Multiagent system on small language models with RAG. Expert-level on conceptual genomics tasks.
Biomni	Stanford University, Genentech, Arc Institute	Biology	ACADEMIA, INDUSTRY, NONPROFIT	General-purpose biomedical agent across 25 fields. (See Chapter 6 for detailed coverage.)

Figure 5.2.8

Earth Science

Progress in AI for Earth science remains aligned with observational infrastructure, including reanalysis datasets and global satellite archives. Weather forecasting, which benefits from decades of reanalysis datasets such as ERA5 and dense global satellite archives, has advanced furthest, with multiple AI models being used in real forecasting systems in 2025. Climate modeling lags behind because it requires projections on decadal timescales where future states fall outside the distribution of any existing training data. Hydrology offers one of the clearest examples of benchmark-driven progress in scientific AI. LSTM-based models, trained jointly across hundreds of catchments in the CAMELS dataset, have consistently outperformed process-based hydrologic models (Kratzert et al., 2019), and regional extensions now span the United States, the United Kingdom, Australia, Chile, and Brazil. Agriculture presents the opposite pattern, with

10 Full references are provided in the Appendix.

shared benchmark datasets still scarce, making progress difficult to measure across research groups despite promising work in knowledge-guided approaches to carbon cycle quantification (Liu et al., 2024) and global change ecology (Jin et al., 2026).

Datasets

Earth science relies heavily on large governmental and institutional observation systems rather than purpose-built AI training corpora. In carbon flux research, global flux tower networks provide foundational observational data. FLUXNET2015 aggregates eddy covariance measurements from sites worldwide (Pastorello et al., 2020), while regional networks including AmeriFlux (North America), ICOS (Europe), and JapanFlux (Japan and East Asia) contribute additional coverage. These datasets enable training and evaluation of models that upscale local carbon flux measurements to regional and global estimates (Figure 5.2.9).

Selected datasets in Earth science (2025)

Name	Affiliation[11]	Domain	Sector	Summary
AmeriFlux	Indiana University, USDA Agricultural Research Service, University of Wisconsin-Madison	Ecology	ACADEMIA, GOVERNMENT	North American network of 260-plus flux tower sites measuring ecosystem carbon, water, and energy exchange. Over 50 sites with 10-plus years of continuous data.
CAMELS	NSF National Center for Atmospheric Research, U.S. Geological Survey, U.S. Department of the Interior	Hydrology	GOVERNMENT	Standardized data on terrain, climate, soil, and streamflow for 671 U.S. river basins. Foundation for AI hydrology benchmarking.
FLUXNET2015	Lawrence Berkeley National Laboratory, University of Tuscia, ETH Zurich	Ecology	ACADEMIA, GOVERNMENT	Global measurements of CO2, water, and energy exchange between ecosystems and atmosphere from 212 sites worldwide.
ICOS	ICOS ERIC/University of Helsinki, Thünen Institute of Climate-Smart Agriculture, ETH Zurich	Ecology	ACADEMIA, GOVERNMENT	European observation network of 140-plus stations across 12 countries measuring greenhouse gas concentrations and carbon fluxes across atmosphere, land, and ocean.
JapanFlux	Osaka Metropolitan University, Chiba University, National Institute of Polar Research (NIPR)	Ecology	ACADEMIA, GOVERNMENT	Land-atmosphere flux measurements covering Japan and East Asia from 1990 to 2023. Tracks energy, water, and CO2 exchange across Asian ecosystems.

Figure 5.2.9

11 Full references are provided in the Appendix.

Benchmarks

In 2025, benchmarks in Earth science expanded into reliability of extreme event coverage, where AI weather models, for example, face the highest stakes. It is also where standard average skill metrics fail to capture performance (Figure 5.2.10).

Selected benchmarks in Earth science (2025)

Name	Affiliation[12]	Domain	Sector	Summary
EarthSE	Shanghai Jiao Tong University, Shanghai Artificial Intelligence Laboratory, Hong Kong Polytechnic University	Earth Science	ACADEMIA, INDUSTRY	100K papers, 114 disciplines, 11 LLMs tested. Significant gaps in Earth science exploration.
ExEBench	Technical University of Munich (TUM), Munich Center for Machine Learning (MCML)	Atmospheric Sciences	ACADEMIA	7 extreme event categories. Global coverage. Tests detection, monitoring, and forecasting.
UnivEARTH	Cornell University, Columbia University	Earth Science	ACADEMIA	140 Earth observation questions. LLM agents: 33% accuracy. Code fails 58% of time.

Figure 5.2.10

Foundation Models

Earth science foundation models released in 2025 covered weather forecasting, climate emulation, and geospatial representation. In weather forecasting, several systems built directly on models highlighted in the 2025 AI Index. FourCastNet 3 generates a 60-day global forecast at 0.25-degree resolution in under 4 minutes on a single GPU, running 8 to 60 times faster than prior approaches. For Earth observation, TerraMind is the first any-to-any generative multimodal model operating across 9 geospatial modalities (Figure 5.2.11).

12 Full references are provided in the Appendix.

Selected foundation models in Earth science (2025)

Name	Affiliation[13]	Domain	Sector	Summary
AlphaEarth	Google DeepMind	Earth observation	INDUSTRY	Embedding field model for general geospatial representation. Outperforms prior featurization approaches.
cBottle (Climate in a Bottle)	Nvidia	Climate Science	INDUSTRY	Diffusion-based climate emulator. Global 5 km at 12.5M-pixel resolution. Diurnal-to-seasonal variability.
FourCastNet 3	Nvidia, Lawrence Berkeley National Laboratory, UC Berkeley	Weather Forecasting	INDUSTRY, GOVERNMENT, ACADEMIA	60-day forecast in <4 min/GPU. 8–60x faster. Builds on Aurora and NeuralGCM advances.
GAIA	USRA/RIACS, BCG X AI Science Institute	Atmospheric Sciences	INDUSTRY, NONPROFIT	Atmospheric FM from 15 years of satellite imagery. Atmospheric rivers (F1: 0.58), cyclone detection (81% recall).
OlmoEarth	Allen Institute for AI, University of Washington, Arizona State University	Earth Observation	NONPROFIT, ACADEMIA	Multimodal spatiotemporal Earth observation FM. Platform for NGOs and nonprofits.
TerraMind	IBM Research, ETH Zurich, Forschungszentrum Jülich	Earth Observation	INDUSTRY, ACADEMIA, GOVERNMENT	First any-to-any generative multimodal Earth observation FM. 9 geospatial modalities.
WeatherNext 2	Google DeepMind	Weather Forecasting	INDUSTRY	Hundreds of weather outcomes in <1 min/TPU. 99.9% improvement over predecessor. Builds on GenCast.

Figure 5.2.11

13 Full references are provided in the Appendix.

AI Agents

In Earth science, AI agents are moving beyond data retrieval toward executing full research workflows, including automated observation processing, literature-informed analysis, and climate task completion. ClimateAgent completed 85 climate tasks with 100% completion and a quality score of 8.32, compared with 6.27 for Microsoft Copilot and 3.26 for GPT-5 (Figure 5.2.12).

Selected AI agents in Earth science (2025)

Name	Affiliation[14]	Domain	Sector	Summary
ClimateAgent	The Hong Kong University of Science and Technology	Climate Science	ACADEMIA	85 climate tasks: 100% completion, quality 8.32 vs. Copilot 6.27, GPT-5 3.26.
EarthLink	Shanghai Artificial Intelligence Laboratory, Fudan University, University of Sydney	Climate Science	INDUSTRY, ACADEMIA	First AI copilot for Earth scientists. Automated research workflows with dynamic feedback loop.
PANGAEA GPT	Alfred Wegener Institute for Polar and Marine Research	Earth Science	GOVERNMENT	Multi-agent system for PANGAEA Earth science database. Intelligent data processing, natural language interface.

Figure 5.2.12

Mathematics

Mathematical reasoning is another active testing ground for AI capabilities. Systems such as Goedel-Prover are moving toward automated formal proof generation in languages like Lean. Competition-level problem-solving and formal verification of known results are advancing quickly, but major open problems, such as long-standing Erdos conjectures, remain well beyond current capabilities. Chapter 2 covers benchmark performance in detail, including a jump from silver to gold medal at the International Mathematical Olympiad in a single year and rapid gains on FrontierMath and MathArena.

14 Full references are provided in the Appendix.

5.3 AI Agents and Tools for Science Workflows

The domain-specific tables of Section 5.2 catalog a growing inventory of AI agents, foundation models, datasets, and benchmark suites. Two cross-domain benchmarks released in 2025 offer a broader view of how well these systems perform when asked to do end-to-end scientific research rather than isolated tasks. On both benchmarks, even the best-performing agents fall well below expert-level performance.

AstaBench

AstaBench is an end-to-end benchmark suite that evaluates agentic scientific research ability across over 2,400 problems spanning multiple domains and the full discovery workflow, from literature understanding through code execution, data analysis, and end-to-end discovery. It benchmarked 57 agents across 22 agent classes and reported both an overall score and cost per problem (Figure 5.3.1). The best performing agent scored around 0.53 at a cost of roughly $3.40 per problem, while most agents clustered between 0.10 and 0.45 at per-problem costs below $1.00.

AstaBench: average score

Source: AstaBench Leaderboard, 2026 | Chart: 2026 AI Index report

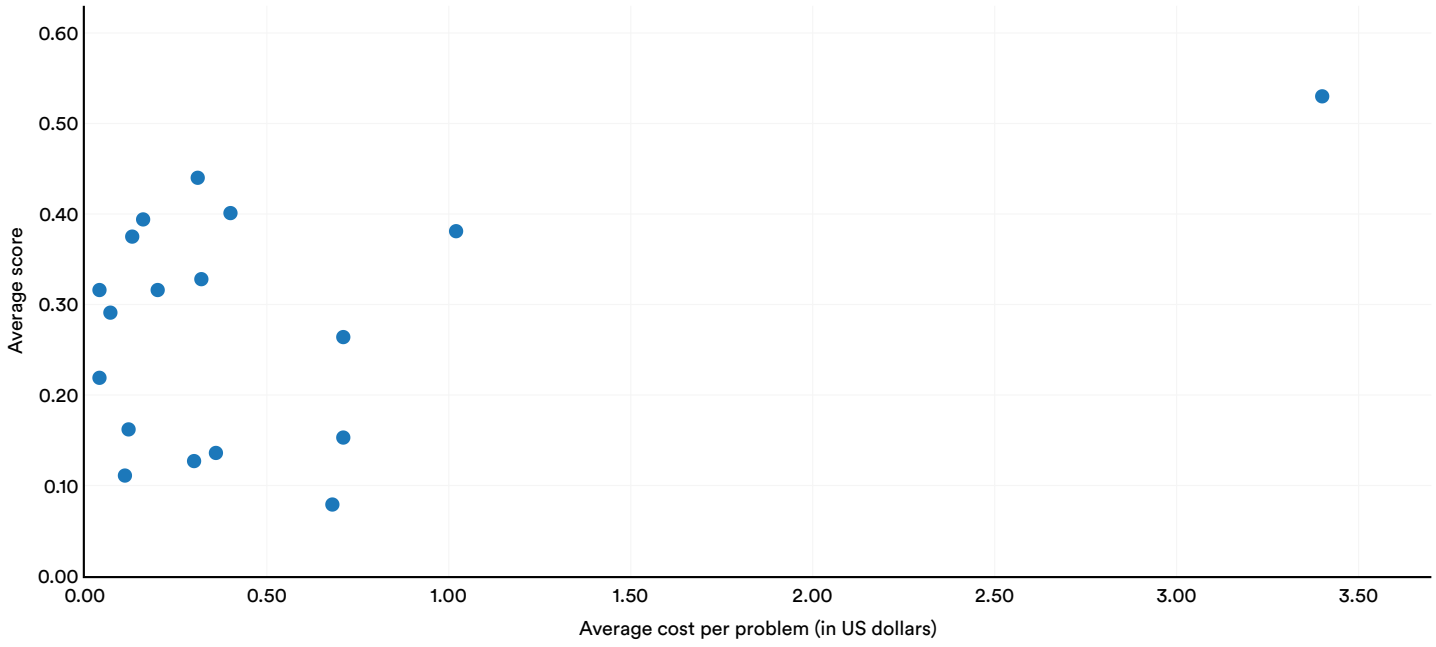


Figure 5.3.1

PaperArena

PaperArena is a benchmark that tests whether LLM agents can answer real research questions that require stitching together evidence across multiple papers while orchestrating external tools for parsing, retrieval, and computation. Gemini 2.5 Pro performs best overall, achieving 38.8% average accuracy in a multiagent configuration (Figure 5.3.2). All tested agents lagged substantially behind the PhD expert baseline of 83.50%. Multiagent configurations consistently outperformed single-agent setups across all models tested, though the gains were modest, typically 2 to 4 percentage points.

PaperArena: single vs. multiagent performance

Source: Wang et al., 2026 | Chart: 2026 AI Index report

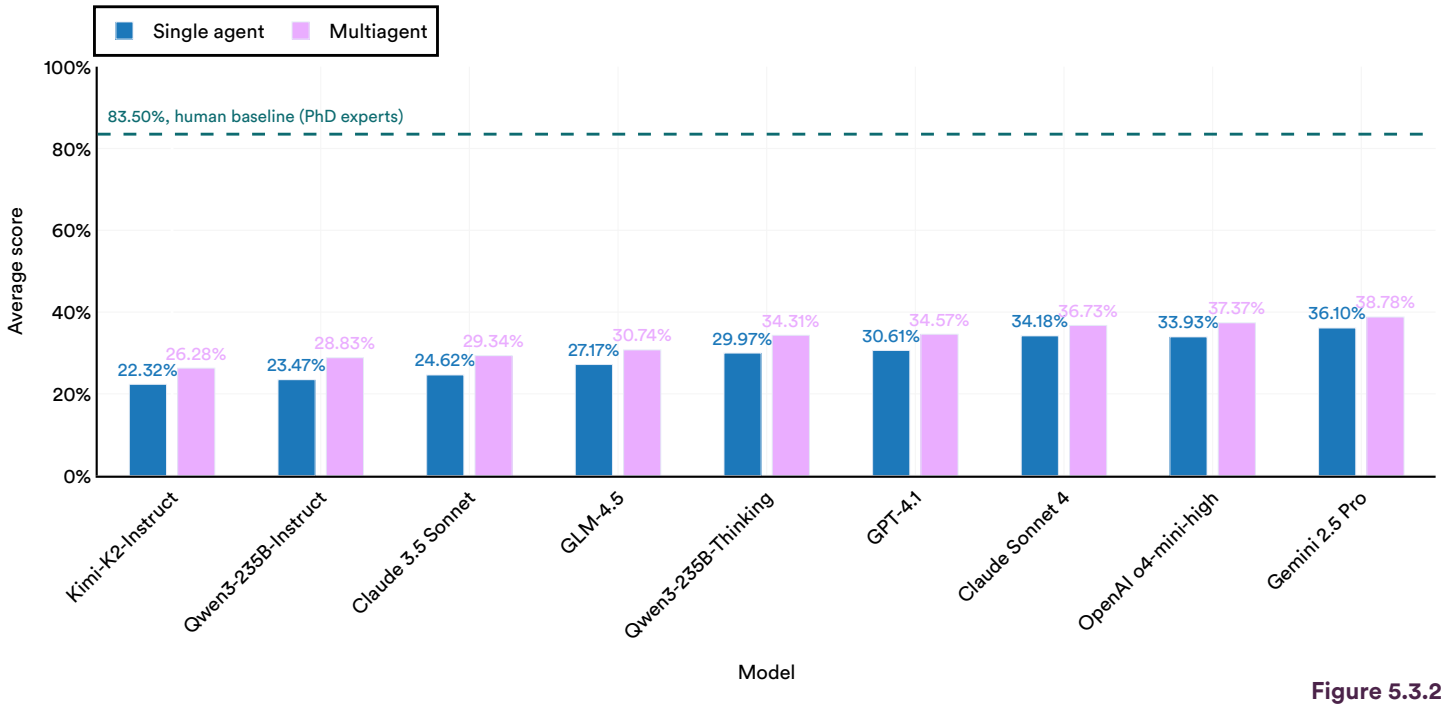


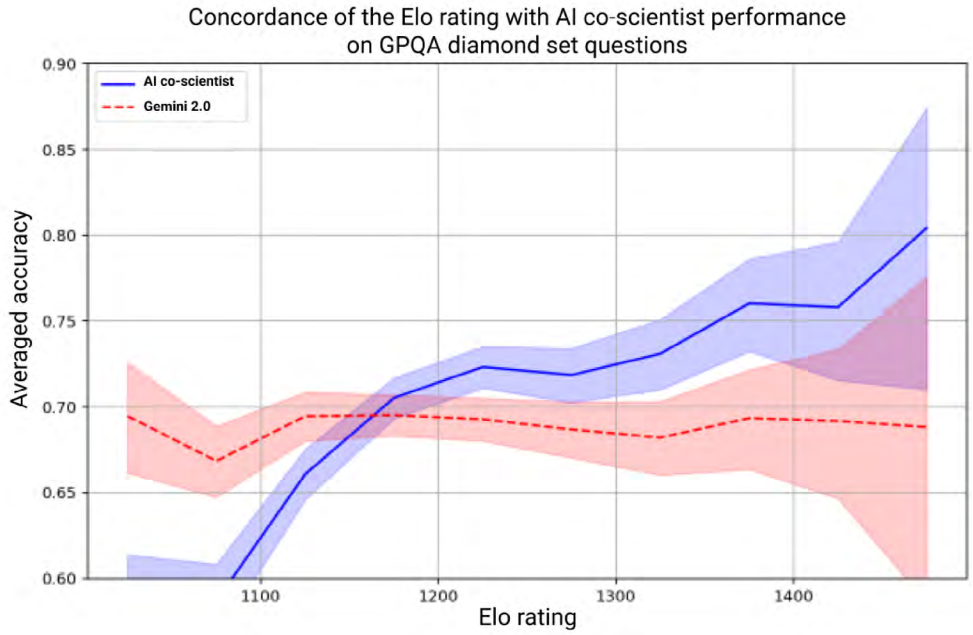
Figure 5.3.2

AI Agents

AI as a Co-scientist

In 2025, several research groups released systems in which multiple AI agents divide scientific tasks among themselves, with separate agents handling literature search, hypothesis generation, code execution, and review. The multiagent systems are designed to approximate the structure of a human research team, rather than relying on a single model or person to perform every step. The most prominent example, Google's AI Co-scientist (Gottweis et al., 2025), uses a generate-debate-evolve loop in which agents iteratively produce and refine evidence-grounded hypotheses. The system was validated in three biomedical areas, including AML drug repurposing and liver fibrosis targets, and achieved a top-1 accuracy of 78.4% on the GPQA Diamond set when selecting its highest-rated hypothesis per question (Figure 5.3.3).

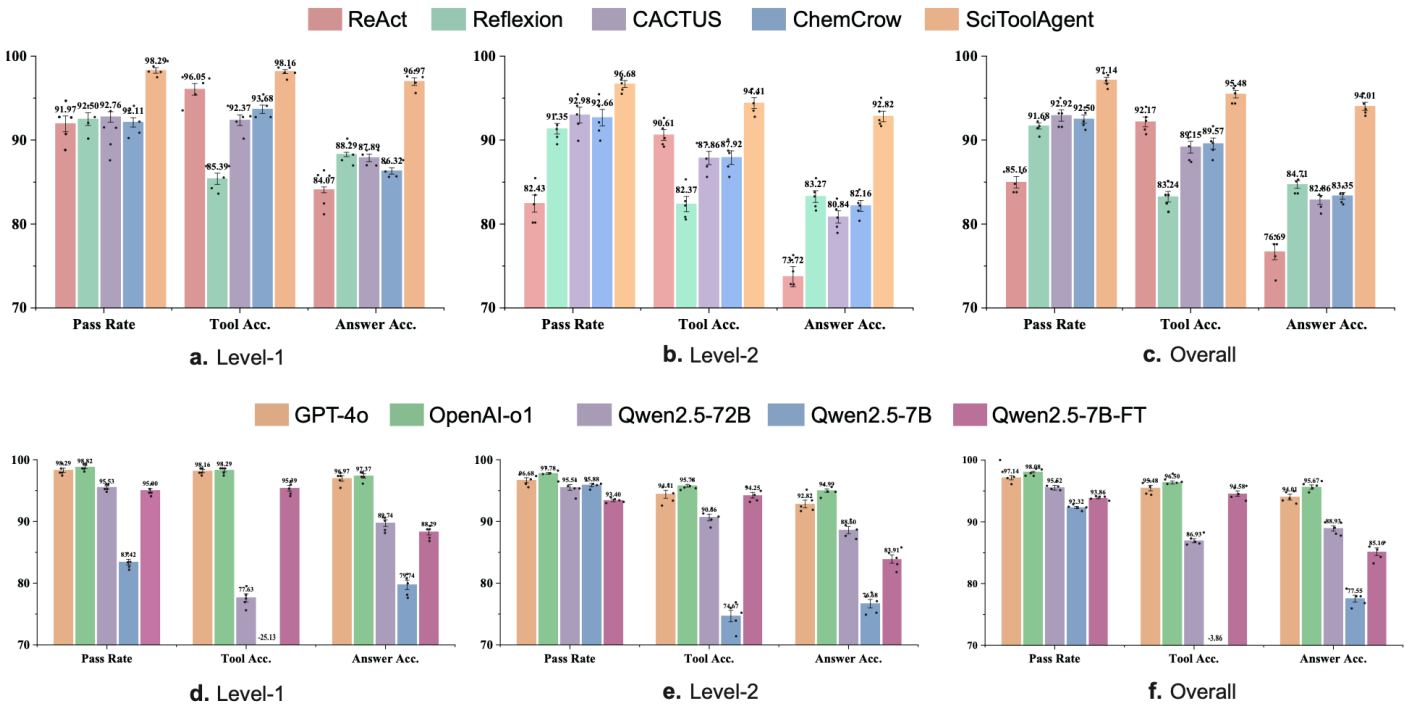
Other multiagent systems pursued different approaches to the same goal. Sakana's AI Scientist-v2 (Yamada et al., 2025) produced the first fully AI-generated paper accepted at a peer-reviewed workshop (ICLR), using agentic tree search to generate and refine code implementations without human-coded templates. Kosmos (Mitchener et al., 2025) maintained coherence across runs lasting up to 12 hours, executing an average of 42,000 lines of code and reading 1,500 papers per run, with collaborators reporting that a single run approximated six months of research. SciToolAgent (Ding et al., 2025) automates hundreds of scientific tools across biology, chemistry, and materials science via knowledge-graph-driven retrieval, outperforming prior agent frameworks by 10 to 20 percentage points on multi-tool tasks (Figure 5.3.4).



Source: Gottweis et al., 2025

Figure 5.3.2

Despite these advances, only a handful of multiagent systems have produced results that were tested and confirmed through real-world experiments. Published examples include new proteins designed by ProtAgents; 92 antibody candidates for SARS-CoV-2 from the Virtual Lab (of which more than 90% successfully bound their target); two new cancer drug targets, GPR160 and ARG2, from OriGene; five novel metal-organic frameworks from MOFGen; and a novel chromophore from ChemCrow. The gap between what these systems can propose computationally and what has been confirmed experimentally remains wide. Key roadblocks for the field include workforce training gaps, a lack of API and interoperability standards, and funding structures that do not yet support the maintenance and scaling of autonomous research infrastructure.



Source: Ding et al., 2025

Figure 5.3.2