OpenDiscovery: A Verifiable, Creative Science Problem-Solving Dataset to Forge AI Scientists

Yixuan Weng[1,†],**Qiyao Sun**[1,†], **Minjun Zhu**[1,2,†], **Yue Zhang**[1*]

[1]School of Engineering, Westlake University [2]Zhejiang University
zhangyue@westlake.edu.cn

Abstract

The grand vision of an "AI Scientist" is constrained by a fundamental bottleneck: while current models excel at imitating known knowledge, they lack the capacity for autonomous discovery and creative problem-solving. We propose OpenDiscovery, a novel dataset paradigm designed to train and benchmark AI agents for verifiable scientific discovery. Moving beyond static input-output pairs, each instance in OpenDiscovery is a self-contained, Dockerized scientific challenge. The AI's task is not merely to predict, but to act—to autonomously formulate a hypothesis, conduct experiments and analysis to arrive at new scientific discoveries and their explanations. The immediate, verifiable feedback from this environment provides an ideal training ground for Reinforcement Learning, aiming to elevate AI from knowledgeable assistants to genuine creative partners in science.

The next frontier in AI for science[1, 2, 3] is to build agents that can collaboratively solve meaningful scientific problems alongside human researchers. A key advantage of such agents is their ability to tirelessly explore vast solution spaces [4]. However, the current development path is hindered by a reliance on complex prompt engineering and manually scripted workflows. This means an agent's "discovery" is predetermined by its human designers, not learned from its own experience [5]. It cannot, for instance, learn from a failed experiment to refine its strategy for the next attempt.

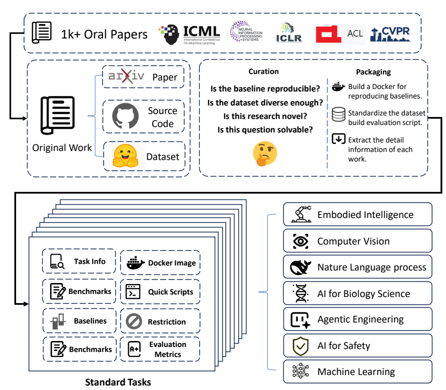

Figure 1: **OpenDiscovery includes 1k of packaged experiment unit.**

This lack of autonomous learning stems from the absence of a proper training environment—a "virtual lab" where agents can engage in large-scale trial-and-error and receive clear, instantaneous feedback. The remarkable success of models like DeepSeek-R1 [6] in mathematics and code [7, 8] highlights a proven principle: progress accelerates when models move from passive supervised learning to active exploration guided by verifiable reward signals. OpenDiscovery aims to systematically apply this principle to the broader domain of scientific discovery.

The vision of an "AI Scientist" [1, 2, 3] is to create an agent that can collaborate with human researchers to tackle grand challenges like climate change and disease treatment. A major advantage of such an agent lies in its ability to tirelessly and fully automatically conduct large-scale exploration and experiments [4], breaking through the inherent limitations of human researchers in time and

*Corresponding Author. Supported by Research Center for Industries of the Future, Westlake University.
† These authors contributed equally to this work, and the order of authors is random.

Published in the AI for Science workshop (NeurIPS 2025)

energy. However, the current path towards this goal has a key bottleneck: when we build so-called "AI Scientist" agents, we often rely on complex, manual prompt engineering and hand-designed execution flows [5]. This means the agent's discovery capability is pre-programmed by human engineers, not learned from its own experience. The LLM-based AI Scientist cannot autonomously learn from a failed experiment to optimize its next discovery strategy. This lack of learning ability stems from the absence of a suitable "training ground"—an environment where an AI agent can conduct large-scale trial-and-error and receive clear, immediate feedback. The immense success of recent models like DeepSeek-R1 [6] in mathematics and code [7, 8] has already shown us the way forward. These models achieved performance beyond supervised imitation learning precisely by constructing reinforcement learning scenarios that rely on verifiable reward signals (e.g., whether a math problem is solved correctly) [9, 10]. The fundamental goal of OpenDiscovery is to systematically apply this proven principle to the broader domain of scientific discovery.

To illustrate this "virtual laboratory," consider protein structure prediction (inspired by AlphaFold [11]). Each OpenDiscovery sample contains four standardized components: 1) A `problem_description.md` specifying the research task, such as "design a more efficient Evoformer module [12] that reduces inference time by 15% while maintaining lDDT accuracy"; 2) A `baseline_code` directory with a runnable baseline model and clearly identified target module (`evoformer_block.py`); 3) A `Dockerfile` providing a reproducible environment with all dependencies (PyTorch, JAX, bioinformatics tools), eliminating setup complexity; 4) A hidden `evaluation_code.py` script containing private test sequences that automatically evaluates submissions and outputs composite performance scores. This structure ensures clear objectives, objective evaluation, and easy reproducibility across research teams.

To ensure the feasibility of this dataset, we have planned a concrete implementation path. Our core strategy is to efficiently transform and leverage existing human intellectual achievements. In the initial phase, we will systematically search existing research papers. For example, we will analyze the approximately 460 papers published in the NeurIPS 2024 Datasets and Benchmarks track, as well as relevant work from other top-tier conferences like ICLR, ACL, and CVPR. We will hire machine learning engineers to manually process these papers and their accompanying datasets and code, carefully packaging the core scientific problems into the standardized format described in the previous section. Throughout this process, we will leverage advanced AI code-assistive tools to improve efficiency, but it will remain a human-expert-led curation process to ensure each problem is of high quality and scientific value.

By establishing a standardized benchmark for creative problem-solving, OpenDiscovery will provide a clear path to accelerate both AI model iteration and practical scientific applications. It will allow the vague concept of "scientific discovery capability" to be quantitatively measured for the first time. This quantitative feedback will greatly facilitate rapid iteration on model architectures and training algorithms (especially RL algorithms). More importantly, the success of OpenDiscovery in computer science will provide an invaluable "proof-of-concept" and a replicable framework. In the future, by collaborating with domain experts, we can replace the verification engine with chemical molecular simulators or gene sequence analysis tools to spawn ‘OpenDiscovery-Chem‘ or ‘OpenDiscovery-Bio‘, extending its empowering potential to nearly all computation-driven scientific disciplines. In summary, our proposal can be distilled into the following three core actions:

- **To construct a reinforcement learning environment centered on verifiable feedback.** We will draw upon the recent successes of AI in mathematics and code, moving beyond passive imitation learning to create a "training ground" for scientific discovery that allows AI to actively explore, trial, and learn from objective rewards.
- **To adopt a pragmatic, efficient, and reproducible dataset curation methodology.** We will build a high-quality, content-rich initial dataset within a reasonable budget and timeframe by systematically transforming published research from top-tier academic conferences and empowering our manual curation process with AI-assistive tools.
- **To build and open-source a standardized "Problem Development Kit".** This toolkit will encapsulate our best practices, providing command-line tools and templates to allow any researcher to easily package their own scientific problem into a standard, auto-evaluable Docker environment. We call for and will support future dataset and benchmark projects to use this toolkit, in order to promote a compatible and interoperable benchmark ecosystem, thereby accelerating the progress of the entire community.

Acknowledgement

We thank the reviewers for their insightful suggestions and valuable feedback. This research is supported by the National Natural Science Foundation of China (NSFC) Key Program under Grant Number 62336006 and 625B2152. Yue Zhang is supported by the Research Program of No. WU2023C020, Research Center for Industries of the Future, Westlake University.

References

[1] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[2] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

[3] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

[4] Qiujie Xie, Yixuan Weng, Minjun Zhu, Fuchen Shen, Shulin Huang, Zhen Lin, Jiahui Zhou, Zilan Mao, Zijie Yang, Linyi Yang, et al. How far are ai scientists from changing the world? *arXiv preprint arXiv:2507.23276*, 2025.

[5] Minjun Zhu, Qiujie Xie, Yixuan Weng, Jian Wu, Zhen Lin, Linyi Yang, and Yue Zhang. Ai scientists fail without strong implementation capability. *arXiv preprint arXiv:2506.01372*, 2025.

[6] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[7] Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, et al. Deepcoder: A fully open-source 14b coder at o3-mini level. *Notion Blog*, 2025.

[8] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

[9] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

[10] Jialong Wu, Shaofeng Yin, Ningya Feng, and Mingsheng Long. Rlvr-world: Training world models with reinforcement learning. *arXiv preprint arXiv:2505.13934*, 2025.

[11] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[12] Ning Sun, Xingyi Cheng, Shentong Mo, Chiming Liu, Hui Li, Eric Xing, and Le Song. Liteformer: Lightweight evoformer for protein structure prediction, 2024.

A Ethical and Safety Considerations

We acknowledge that an AI system capable of autonomous discovery carries inherent dual-use risks. A model trained to discover beneficial molecular structures or efficient algorithms could theoretically be applied to harmful ends, such as designing novel toxins or cyber-attacks. Therefore, a proactive and rigorous ethical framework is a cornerstone of the OpenDiscovery project, not an afterthought. Our primary mitigation strategy is a strict, domain-level exclusion policy implemented during the curation process. We will actively avoid sourcing problems from domains with clear potential for misuse or societal harm. This includes, but is not limited to, weapons development, the synthesis of toxic chemicals or pathogens, and the creation of novel offensive cybersecurity exploits. Our focus will be intentionally directed towards problems with clear, positive societal value, such as improving computational efficiency, advancing fundamental scientific understanding, and solving challenges in areas like sustainable energy or medical diagnostics.

Furthermore, our initial data sourcing methodology provides a powerful, built-in ethical safeguard. In the project's early phase, our feasibility relies heavily on adapting published papers from top-tier conferences like NeurIPS, ICLR, ACL, and CVPR. These conferences have mature and stringent ethical review processes; any accepted paper, especially within the datasets and benchmarks tracks, must have already addressed its potential ethical and societal impacts. By curating from this pre-vetted pool of research, our initial dataset largely inherits the robust ethical assurances established by the academic community, significantly minimizing the risk of incorporating harmful or controversial problems from the outset. Looking forward, as OpenDiscovery transitions to a community-driven model, we will establish an Ethical Review Board. This committee, comprised of experts in AI, ethics, and relevant scientific domains, will be responsible for reviewing all community-submitted problems via the Problem Development Kit (PDK) to ensure they align with our safety guidelines before being integrated into the official benchmark.

B Broader Impacts and Community Discussion

The advent of AI Scientists, trained on dataset like OpenDiscovery, will likely catalyze a profound shift in the role of the human researcher. As AI agents become increasingly capable of handling the exhaustive and iterative aspects of exploration and optimization, the premium on human intellect will move towards higher-level cognitive tasks. We anticipate that the most critical role for human researchers will be to formulate clear, scientifically significant, and ethically sound research questions. The creative spark to identify a novel problem or a promising, unexplored direction will remain a fundamentally human contribution. Subsequently, the detailed, methodical development and search for a solution within that defined scope can be delegated to the AI Scientist. This fosters a new paradigm of human-AI collaboration, where the human acts as the "Principal Investigator," providing strategic guidance and high-level direction—such as suggesting novel approaches or adjusting research priorities based on intermediate results—rather than letting the AI engage in completely unconstrained exploration. This partnership leverages the AI's tireless, high-throughput capabilities while keeping human intuition and judgment at the helm of the scientific endeavor.

Furthermore, the process of training these AI Scientists via reinforcement learning necessitates new modes of supervision and safety protocols. The learning process itself, during which the AI agent explores countless potential solutions, must be carefully contained to prevent unintended consequences. Our OpenDiscovery framework is designed with this in mind. The strict sandboxing provided by the Docker environment is a critical safety feature. During training and evaluation, each container can be configured to operate in a highly restricted mode, most importantly with external network connections completely disabled. This crucial step ensures that the AI agent's actions are strictly confined to the problem environment. It is prevented from accessing external APIs, downloading arbitrary code, or interacting with other systems, thereby mitigating the risk of unintentional harm or information leakage. This principle of "contained experimentation" is fundamental to the responsible development of autonomous discovery agents, allowing the research community to harness the benefits of their tireless exploration while maintaining robust control and safety over the process.