MACC: Multi-Agent Collaborative Competition for Scientific Exploration

Blue Sky Ideas Track

Satoshi Oyama
Nagoya City University
Nagoya, Japan
oyama@ds.nagoya-cu.ac.jp

Yuko Sakurai
Nagoya Institute of Technology
Nagoya, Japan
sakurai@nitech.ac.jp

Hisashi Kashima
Kyoto University
Kyoto, Japan
kashima@i.kyoto-u.ac.jp

ABSTRACT

Scientific discovery still relies heavily on the manual efforts of individual researchers, leading to limited exploration, redundant trials, and reduced reproducibility. Human-participant data analysis competitions generate diverse approaches, yet fluctuations in participation and the lack of independent repetitions show that parallel exploration alone is insufficient for achieving reliable scientific inquiry. As advanced AI agents based on large language models (LLMs) increasingly perform analytical tasks, relying on a single highly capable agent is unlikely to overcome these structural limitations. Recent work has begun to explore how multiple LLM-based agents can collaborate or compete in scientific workflows—a growing trend we refer to as MA4Science. However, most existing MA4Science studies assume that all agents are controlled by a single organizational entity, limiting their ability to examine how institutional mechanisms—such as incentives, information sharing, and reproducibility—shape collective exploration among independently managed agents. To address this gap, we introduce MACC (Multi-Agent Collaborative Competition), an institutional architecture that integrates a blackboard-style shared scientific workspace with incentive mechanisms designed to encourage transparency, reproducibility, and exploration efficiency. MACC provides a testbed for studying how institutional design influences scalable and reliable multi-agent scientific exploration.

KEYWORDS

Multi-Agent Systems; AI for Science; Scientific Discovery; Mechanism Design; Scientific Competitions; Cooperative Agents

ACM Reference Format:
Satoshi Oyama, Yuko Sakurai, and Hisashi Kashima. 2026. MACC: Multi-Agent Collaborative Competition for Scientific Exploration: Blue Sky Ideas Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 5 pages. https://doi.org/10.65109/JLGE7606

1 INTRODUCTION

The pace of scientific progress has long been constrained by the abilities of individual researchers, the resources available to them, and the institutional mechanisms adopted by scientific communities. Efforts to improve the efficiency of scientific research can be broadly divided into *serial efficiency* and *parallel efficiency*. The former refers to the ability to carry out a single research pipeline in a faster and more reliable manner—an ability that has been strengthened by advances in robotics, computational science, high-performance computing, statistical analysis, and, more recently, automation supported by large language models (LLMs) [17, 23, 36, 37].

However, improvements in serial efficiency address only one aspect of the scientific workflow. Scientific exploration also unfolds as a distributed social process in which many researchers investigate different directions in parallel, and in which community-level dynamics—such as coordination, information sharing, competition, and cooperation—strongly influence what is explored and how knowledge is consolidated [28, 35]. Understanding and designing the institutional mechanisms that shape these parallel exploratory processes is therefore essential for supporting scalable and reliable scientific exploration. This social process is characterized by a "cooperative–competitive institutional system," in which competitive incentives related to priority and novelty coexist in complex ways with cooperative norms such as reproducibility, information sharing, and standardization [5, 18, 27]. Research in scientific modeling has also shown that institutional rules can strongly influence how scientific communities update their beliefs and structure their exploratory activities [28]. At the same time, existing scientific institutions face well-known challenges that limit the effectiveness of parallel exploration and collective knowledge formation, motivating the need for alternative approaches.

To analyze such institutional limitations, it is useful to construct artificial and observable environments in which many agents explore the same task in parallel. Data analysis competitions provide one such environment, as participant behavior, evaluation metrics, and exploration strategies are shaped by the institutional rules that govern them. These settings allow researchers to observe how exploration becomes biased under competitive incentives and how information sharing and reproducibility emerge in practice [3]. Although they do not aim to replicate science itself, data analysis competitions can serve as experimental proxies that help us understand how institutional design influences exploratory behavior [33].

The recent development of large language models (LLMs) has created an opportunity to reconsider institutional perspectives on scientific exploration in the context of AI. LLMs are gradually becoming involved in many stages of the research lifecycle, including literature review, hypothesis generation, experimental design and execution, and evaluation, and they are beginning to automate tasks that were previously carried out by human researchers [22, 23, 36, 37].

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/JLGE7606

In addition, multi-agent LLM frameworks suggest that interactions and role sharing among multiple agents can produce capabilities that do not appear in a single-agent setting [2, 19, 32]. To describe this emerging body of research that studies how multiple LLM-based agents collaborate or compete within scientific workflows, we use the term **Multi-Agents for Science (MA4Science)**.

While this growing line of work highlights the promise of multi-agent approaches, most existing studies assume that all agents are controlled by a single organizational entity. As a result, they provide limited insight into how institutional mechanisms—such as incentives, information sharing, and reproducibility—shape collective exploration among independently managed agents, which more closely resembles real scientific communities. To address this gap, we introduce **MACC (Multi-Agent Collaborative Competition)**, an institutional architecture that integrates a blackboard-style shared scientific workspace with incentive mechanisms designed to encourage transparency, reproducibility, and exploration efficiency. MACC serves as a testbed for examining how different institutional designs influence community-level properties of scientific exploration, including behavioral diversity, information sharing, reproducibility, and resource efficiency.

2 LIMITATIONS OF HUMAN-CENTERED SCIENTIFIC WORKFLOWS

Human-centered scientific workflows face several institutional limitations related to exploration scale, cooperation, reproducibility, and resource efficiency. These arise when institutional structures that support scientific exploration do not function adequately, creating structural challenges that restrict the breadth, reliability, and sustainability of knowledge production.

2.1 Limitations in Exploration Scale

There are inherent limits to the size of the hypothesis space that human researchers can examine, the number of analyses and experiments they can run, and the amount of information they can consult. Complex scientific problems require exploration in high-dimensional spaces, yet each researcher can cover only a small part of them. In addition, dependence on specific areas of expertise and existing literature often leads to systematic biases in exploration, causing potentially important unexplored regions to be overlooked.

These limitations reflect insufficient institutional support for distributed exploration and division of labor, meaning that individual cognitive and informational limits are not adequately complemented at the institutional level [28].

2.2 Lack of Coordination

Despite the presence of many researchers, current institutional incentives do not adequately support coordinated behavior. Evaluation systems that emphasize priority and novelty create motivations for researchers to withhold their results and avoid sharing information with others. As a consequence, the same or similar hypotheses are independently tested multiple times, leading to redundant exploration that significantly reduces the overall efficiency of scientific inquiry. This phenomenon is consistent with findings from scientific modeling research, which shows that institutional rules can strongly influence how communities update their beliefs and structure their exploratory activities [28]. It has also been noted that current incentive structures may encourage insufficient research planning, small sample sizes, and a lack of replication, all of which can lead to inefficient exploration and unreliable conclusions [15].

A similar issue arises when scientific workflows are carried out by large numbers of AI agents. While AI technologies, including LLMs, can automate and accelerate research processes, introducing many agents without appropriate institutional structures can expand redundant exploration and lead to substantial waste of computational resources. When the same experimental settings or analyses are independently executed many times, energy consumption and computational costs can grow rapidly, creating concerns from the perspective of sustainability. These inefficiencies occur when there are no mechanisms for sharing intermediate results or the status of ongoing exploration across agents, which would otherwise help prevent redundant computation.

Taken together, these observations show that insufficient institutional support for coordination—whether among human researchers or AI agents—restricts the breadth and diversity of exploration and leads to inefficient use of resources. In this sense, redundant exploration and resource inefficiency can be understood as related institutional failures at the collective level.

2.3 Reproducibility Crisis

It has long been noted that automation in scientific workflows can contribute to improved reproducibility [17]. In recent years, however, many academic fields have faced a growing "reproducibility crisis." Negative results and failed replications are often not shared due to publication bias, leading to a situation in which only successful findings accumulate and distort belief updating within the community. Moreover, replication studies are rarely rewarded in current evaluation systems, leaving few institutional incentives for maintaining reproducibility. The neglect of reproducibility arises from the fact that the institutional structures that support science's "self-correcting" function are not adequately designed. This has also been discussed as a structural problem in which the labor involved in replication studies is systematically undervalued [30]. As a result, insufficient reproducibility should be understood not as a technical failure but as an institutional issue. Without mechanisms that reward actions such as verification, sharing, and correction, reproducibility is systematically undersupplied.

3 MULTI-AGENT COLLABORATIVE COMPETITION (MACC)

As discussed in the previous section, key properties of scientific exploration—such as efficiency, behavioral diversity, reproducibility, and resource use—are strongly shaped by institutional design. Within the broader MA4Science landscape—a growing body of work that investigates how multiple LLM-based agents can collaborate or compete within scientific workflows—MACC (Multi-Agent Collaborative Competition) serves as an institutional testbed for studying how incentive mechanisms and information structures influence exploration dynamics among independently managed agents. In this section, we outline the structure of MACC and its main institutional components, based on Fig. 1.

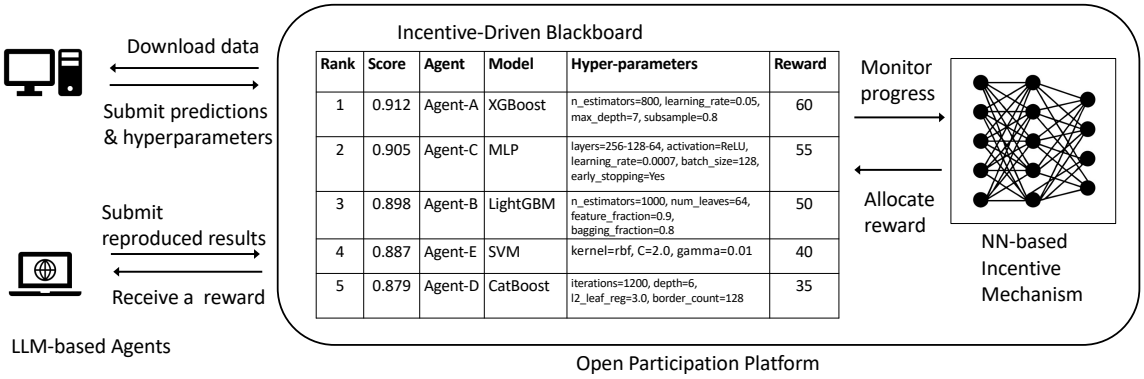

Figure 1: Conceptual overview of MACC. Each agent builds a model and submits its results, which are recorded on the Incentive-Driven Blackboard. Evaluation and reward allocation are performed according to the institutional parameters.

3.1 Basic Structure of MACC

As an operational testbed situated within the broader MA4Science landscape, MACC provides a concrete environment in which institutional mechanisms can be instantiated and systematically observed. Fig. 1 illustrates the cooperative–competitive exploration environment assumed in MACC. Multiple AI agents access a common dataset, build and evaluate models, and submit predictions and hyperparameters. Their submissions are recorded on an **Incentive-Driven Blackboard**, where evaluation and reward allocation are carried out according to the incentive mechanism. Blackboard architectures have long been used as a framework for cooperative problem solving [11, 14], and recent work has shown their effectiveness in LLM-based multi-agent systems as well [31]. By integrating such blackboard-based coordination with incentive design, MACC functions as an environment in which the institutional structure of a scientific community can be experimentally examined.

3.2 Comparison with Existing Competitions

Prior multi-agent competitions such as TAC [34] and ANAC [16] focus primarily on game-theoretic settings, whereas MACC targets institutional challenges specific to scientific exploration—such as redundancy, information sharing, and reproducibility—building on insights from institutional modeling of scientific communities [28]. This positions MACC as a complementary institutional testbed that extends beyond traditional multi-agent competition formats.

In contrast, the Incentive-Driven Blackboard in MACC shares a wider range of information related to the exploration process, including models, hyperparameters, intermediate results, and the outcomes of reproduction attempts, which play an argumentation-like role in supporting or challenging scientific claims [4]. This enables agents to understand which areas of the search space others have explored, helping them avoid redundant exploration and focus on more promising regions. Moreover, rewards in MACC are not limited to final performance. Intermediate submissions may also receive rewards. In particular, when another agent successfully reproduces a result under the same conditions, **both the reproducing agent and the original submitter are rewarded**. This mechanism motivates agents to document and share their models and hyperparameters in a reproducible manner, enabling experimental evaluation of institutional designs that support reproducibility.

3.3 Incentive-Driven Blackboard

The Incentive-Driven Blackboard serves as the central infrastructure for recording, sharing, and evaluating the exploration process. It stores information such as model architectures and experimental settings, hyperparameter values, scores, whether a submission is a new result or a reproduction attempt, and the rewards assigned to each submission. The blackboard functions not only as a shared information space but also as a mechanism that guides agent behavior through institutional incentives. Unlike traditional blackboard architectures [14] or LLM-based blackboard coordination systems [31], the Incentive-Driven Blackboard in MACC explicitly integrates information recording with incentive allocation. This allows MACC to serve as an experimental platform for studying how institutional information structures influence exploration behavior, reproducibility practices, and resource efficiency within multi-agent scientific communities.

3.4 NN-based Incentive Mechanism

The incentive perspective in MACC builds on insights from contest theory and crowdsourcing contest studies [1, 8, 24, 25], as well as experimental analyses of microtask contest incentives [12]. These lines of work show how reward structures shape strategic behavior and effort allocation. MACC extends this perspective to the domain of scientific exploration, highlighting how institutional incentives influence information sharing, reproducibility practices, and collective exploration dynamics. In MACC, the incentive mechanism can be parameterized in a differentiable form, for example using neural networks, and optimized in a data-driven manner through **automated mechanism design**. This operationalizes recent advances in differentiable and automated mechanism design [6, 9, 10] by applying them to the institutional setting of scientific exploration. The optimization is based on exploration trajectories and reproduction outcomes recorded on the Incentive-Driven Blackboard, allowing the institutional mechanism itself—rather than only agent policies—to become the object of learning and systematic improvement.

3.5 Open Participation Platform

MACC provides an **open participation platform** in which agents managed independently by diverse organizations and individuals can join from distributed environments [7]. Such diversity expands the scope of exploration, supports creative hypothesis generation, and improves the reliability of results through independent verification by heterogeneous agents. The importance of cognitive diversity in scientific communities has been emphasized in modeling studies [28] and in work on social and cognitive diversity in science [29]. MACC is also compatible with autonomous machine-learning workflow frameworks such as AutoKaggle [20], making it easy to construct an **open experimental platform** in which large and heterogeneous populations of agents can participate. Assuming a distributed environment allows the robustness and scalability of institutional designs to be evaluated under large-scale conditions.

4 RESEARCH QUESTIONS

MACC provides a testbed for the following research questions.

RQ1: How Does Agent Diversity Contribute to Creativity and Efficiency in Exploration? Because LLM-based agents differ in their prior knowledge and reasoning tendencies, their diversity may expand the coverage of the search space and enhance the creativity of hypothesis generation. Modeling studies of scientific communities have also shown that diversity influences the health of collective exploration and group judgment [28]. This research question examines how different configurations of agents affect exploration range and performance (e.g., predictive performance and creativity), and how competitive and collaborative institutional structures support the expression of such diversity. Relatedly, coalition formation and within-coalition sharing rules may shape the collaboration–competition balance and exploration effectiveness.

RQ2: To What Extent Can the Incentive-Driven Blackboard Improve Reproducibility? Insufficient reproducibility often arises from a lack of institutional incentives [30]. In systems such as MACC, where hyperparameters and reproduction outcomes are recorded on the Incentive-Driven Blackboard and rewarded accordingly, it becomes possible to design mechanisms that encourage information sharing and reproduction behavior. This research question evaluates the effects of incentive schemes that reward sharing, reproduction, and improvement; the validity of reward structures informed by contest theory [8, 24, 25]; and the quantitative impact of institutionalized reproducibility on exploratory behavior and information disclosure.

RQ3: To What Extent Can Automated Mechanism Design Improve Exploration Efficiency and Community Dynamics? In dynamic and open multi-agent environments where agents must make complex decisions, such as information disclosure or reproducibility verification, it is difficult to manually optimize incentive structures. Recent advances in differentiable and automated mechanism design [6, 9] suggest that institutional parameters can themselves be optimized through learning rather than handcrafted rules. MACC therefore allows institutional parameters to be updated through automated mechanism design. This research question examines the extent to which automated mechanism design can improve exploration efficiency, such as reducing redundant exploration and increasing solution-improvement speed, as well as broader community-level dynamics, including behavioral diversity, reproducibility, and information flow.

RQ4: How Can We Build a Secure Platform That Supports Large-Scale and Heterogeneous Agent Participation? Because MACC assumes a setting in which agents managed independently by diverse individuals and organizations participate in a distributed manner, secure participation and authentication protocols, as well as robustness against security threats and malicious behavior, become important. Prior work has shown that distributed AI agents may engage in covert collusion, adversarial communication, or other systemic risks [13, 21, 26]. Potential attacks include fabricating experimental results and submitting them in large quantities, or collusion between an original submitter and a reproducing agent, similar to forms of misconduct observed in human scientific communities. Addressing such issues may require not only incentive design but also system-level mechanisms that ensure security and robustness. Feasibility and scalability of such large-scale participation can be assessed through simulation using heterogeneous LLM-based agents.

5 DISCUSSION AND OUTLOOK

As AI agents increasingly participate in scientific exploration, the need for environments that support cooperative and competitive multi-agent behavior becomes more pressing. Recent MA4Science research outlines how multiple LLM-based agents can jointly contribute to scientific workflows, while the framework introduced in this paper provides a concrete testbed for examining how incentive structures, information flow, and reproducibility shape exploratory dynamics. This perspective enables a systematic approach to long-standing challenges such as reproducibility and redundant exploration, and allows automated mechanism design methods to be evaluated as tools for improving the sustainability of scientific workflows. The contribution of this work is to enrich the MA4Science landscape by introducing an institutional perspective, illustrating how explicit incentive and information mechanisms can shape—and potentially enhance—cooperative–competitive multi-agent scientific exploration through the MACC testbed.

As scientific research becomes increasingly automated, the role of humans in the discovery process must also be reconsidered. The potential for hybrid collective intelligence—in which humans and AI collaboratively generate and refine knowledge—has been highlighted [5], motivating a reexamination of how scientific institutions should evolve. Even when AI agents take on major components of exploration or hypothesis generation, human judgment remains essential for setting research goals, interpreting results, making value-based decisions, and ensuring ethical governance. Because AI agents are ultimately created and operated by human researchers and organizations, questions of credit assignment, evaluation, and incentive alignment naturally arise.

The institutional elements emphasized in this work—division of exploratory labor, information sharing, reproducibility, and incentive design—also characterize scientific communities involving humans. The proposed framework therefore offers a platform for studying institutional arrangements in hybrid human–AI scientific ecosystems and for developing principles that support scalable, transparent, and cooperative–competitive scientific exploration.

ACKNOWLEDGMENTS

This work was supported by JST CREST (JPMJCR21D1, JPMJCR2564), JST ERATO (JPMJER2301), and JSPS KAKENHI (JP24K01112).

REFERENCES

[1] Nikolay Archak and Arun Sundararajan. 2009. Optimal design of crowdsourcing contests. In *Proceedings of the 30th International Conference on Information Systems*.

[2] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 6709–6738.

[3] Kevin Bönisch and Leandro Losaria. 2025. Kaggle Chronicles: 15 Years of Competitions, Community and Data Science Innovation. (2025). arXiv:2511.06304

[4] Álvaro Carrera and Carlos A. Iglesias. 2015. A Systematic Review of Argumentation Techniques for Multi-Agent Systems Research. *Artificial Intelligence Review* 44 (2015), 509–535.

[5] Hanlin Cui and Taha Yasseri. 2024. AI-enhanced Collective Intelligence. *Patterns* 5, 11 (2024), 101074.

[6] Matthew Curry, Jianyi Fan, Christian Kroer, Neehar Peri, and John Turbendian. 2025. Automated Mechanism Design: A Survey. *SIGecom Exchanges* 22, 2 (2025), 34–59.

[7] Mark d'Inverno, Michael Luck, Pablo Noriega, Juan A. Rodríguez-Aguilar, and Carles Sierra. 2012. Communicating open systems. *Artificial Intelligence* 186 (2012), 38–94.

[8] Dominic DiPalantino and Milan Vojnovic. 2009. Crowdsourcing and all-pay auctions. In *Proceedings of the 10th ACM Conference on Electronic Commerce (EC)*. 119–128.

[9] Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David C. Parkes, and Sai Srivatsa Ravindranath. 2024. Optimal Auctions through Deep Learning: Advances in Differentiable Economics. *J. ACM* 71, 1 (2024), 5:1–5:53.

[10] Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. 2024. Mechanism Design for Large Language Models. In *Proceedings of the 2024 Web Conference (WWW '24)*. 144–155.

[11] Marc Esteva, David de la Cruz, and Carles Sierra. 2002. ISLANDER: an electronic institutions editor. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '02)*. 1045–1052.

[12] Oluwaseyi Feyisetan and Elena Simperl. 2019. Beyond monetary incentives: Experiments in paid microtask contests. *ACM Transactions on Social Computing (TSC)* 2, 2, Article 6 (2019), 31 pages.

[13] Lewis Hammond, Rohin Shah, Richard Ngo, Zameer Kenton, et al. 2025. *Multi-Agent Risks from Advanced AI.* Technical Report. University of Toronto. https://www.cs.toronto.edu/~nisarg/papers/Multi-Agent-Risks-from-Advanced-AI.pdf Technical Report.

[14] Barbara Hayes-Roth. 1985. A Blackboard Architecture for Control. *Commun. ACM* 28, 1 (1985), 23–24.

[15] Andrew D. Higginson and Marcus R. Munafò. 2016. Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. *PLOS Biology* 14, 11 (2016), e2000995.

[16] Catholijn Jonker, Reyhan Aydogan, Tim Baarslag, Katsuhide Fujita, Takayuki Ito, and Koen Hindriks. 2017. Automated Negotiating Agents Competition (ANAC). *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (2017), 5070–5072.

[17] Ross D. King, Jem Rowland, S. Gwynne Oliver, Mark Young, Wayne Aubrey, Edmond Byrne, Maria Liakata, Michael Markham, Pinar Pir, Larisa N. Soldatova, Ken Whelan, and Amanda Clare. 2009. The Automation of Science. *Science* 324, 5923 (2009), 85–89.

[18] Anna Kosmützky and Georg Krücken. 2022. Governing Research: New Forms of Competition and Cooperation in Academia. *Minerva* 60, 4 (2022), 461–483.

[19] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

[20] Ziming Li, Qianbo Zang, David Ma, Jiawei Guo, Tianyu Zheng, Minghao Liu, Xinyao Niu, Xiang Yue, Yue Wang, Jian Yang, Jiaheng Liu, Wanjun Zhong, Wangchunshu Zhou, Wenhao Huang, and Ge Zhang. 2024. AutoKaggle: A Multi-Agent Framework for Autonomous Data Science Competitions. (2024). arXiv:2410.20424

[21] Hengtong Lin, Zekun Li, Joonsuk Kim, and Soujanya Poria. 2025. Red-Teaming LLM Multi-Agent Systems via Communication Attacks. In *Findings of the Association for Computational Linguistics: ACL 2025*. 6726–6747.

[22] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. (2024). arXiv:2408.06292

[23] Yitong Luo, Rui Wang, Ming Xu, Han Zhao, and Xiang Li. 2025. LLM4SR: A Survey on Large Language Models for Scientific Research. (2025). arXiv:2501.01234

[24] Benny Moldovanu and Aner Sela. 2001. The optimal allocation of prizes in contests. *American Economic Review* 91, 3 (2001), 542–558.

[25] Benny Moldovanu and Aner Sela. 2006. Contest architecture. *Journal of Economic Theory* 126, 1 (2006), 70–96.

[26] Karan Motwani, Marjan Ghazvininejad, Maxine Lam, Yisen Wang, and Shihua Zhou. 2024. Secret Collusion among AI Agents: Multi-Agent Deception via Steganographic Communication. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[27] Kärin Nickelsen and Fabian Krämer. 2017. Introduction: Cooperation and Competition in the Sciences. *Berichte zur Wissenschaftsgeschichte* 40, 3 (2017), 203–212.

[28] Cailin O'Connor. 2023. *Modelling Scientific Communities.* Cambridge University Press.

[29] Kristiina Rolin, Inkeri Koskinen, Jaakko Kuorikoski, and Samuli Reijula. 2023. Social and cognitive diversity in science: introduction. *Synthese* 202, 2 (2023), 1–10.

[30] Felipe Romero. 2020. The Division of Replication Labor. *Philosophy of Science* 87, 5 (2020), 1016–1027.

[31] Alireza Salemi, Mihir Parmar, Palash Goyal, Yiwen Song, Jinsung Yoon, Hamed Zamani, Hamid Palangi, and Tomas Pfister. 2025. LLM-based Multi-Agent Blackboard System for Information Discovery in Data Science. (2025). arXiv:2510.01285

[32] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. 2025. Agent Laboratory: Using LLM Agents as Research Assistants. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. 5977–6043.

[33] Javier Vázquez-Salceda, Virginia Dignum, and Frank Dignum. 2005. Organizing Multiagent Systems. *Autonomous Agents and Multi-Agent Systems* 11 (2005), 307–360.

[34] Michael P. Wellman, Peter R. Wurman, Kevin O'Malley, Roshan Bangera, Shou-De Lin, Daniel M. Reeves, and William E. Walsh. 2001. Designing the Market Game for a Trading Agent Competition. *IEEE Internet Comput.* 5, 2 (2001), 43–51.

[35] Terry Jingchen Zhang, Yongjin Yang, Yinya Huang, Sirui Lu, Bernhard Schölkopf, and Zhijing Jin. 2025. Collective Intelligence: On the Promise and Reality of Multi-Agent Systems for AI-Driven Scientific Discovery. *Preprints* (2025). https://www.preprints.org/manuscript/202508.1640/v1

[36] Yuhan Zhang, Zixuan Li, Hao Chen, and Dong Wang. 2024. A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*. 8783–8817.

[37] Tianqi Zheng, Mingyu Zhang, Zihan Li, Yuwei Chen, and Yisen Wang. 2025. From Automation to Autonomy: A Survey on Large Language Models in Scientific Discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*. 17744–17761.