Scientific Discoveries by LLM Agents

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large Language Models (LLMs) have evolved from text generators into sophisticated autonomous agents capable of conducting independent scientific research. This paper reviews the current landscape of LLM-driven scientific discovery, where AI agents can now execute the entire research pipeline, including reading scientific literature, forming novel hypotheses, designing experiments, interfacing with laboratory tools and simulators, analyzing data, and interpreting results. A key advancement is the deployment of multi-agent systems, where specialized agents collaborate in roles such as 'scientist,' 'critic,' and 'evaluator' to tackle complex challenges beyond the scope of individual agents. We survey domain-specific applications and highlight validated discoveries, including the autonomous synthesis of novel chemical compounds and materials, the design of functional nanobodies for SARS-CoV-2 variants, and the automation of complex bioinformatics analyses. The development of end-to-end research systems that can progress from an initial idea to a full, peer-reviewed publication demonstrates a paradigm shift in the automation of science. Despite these successes, significant challenges remain, including performance degradation on highly complex causal reasoning tasks. Future directions point toward creating more robust, causally-aware agents and enhancing human-AI collaboration to accelerate scientific breakthroughs.

1 Introduction

Large Language Models (LLMs) are being used as autonomous agents to make real scientific discoveries by reading papers, forming hypotheses, designing experiments, and analyzing results. These AI systems can now work independently or in teams to advance research across many scientific fields.

This paper provides a comprehensive overview of the current state and future potential of LLM agents in scientific discovery. We begin by examining the foundational Capabilities and Core Functions that enable individual LLM agents to mirror the traditional scientific method. From there, we explore the evolution toward collaborative Multi-Agent Systems and Frameworks, where specialized agents work in concert to solve complex problems.

To illustrate these concepts, we survey a range of Domain-Specific Applications, highlighting validated breakthroughs in biology, chemistry, materials science, and healthcare. We then discuss the culmination of this research in End-to-End Autonomous Research Systems capable of managing the entire scientific workflow from initial idea to final publication. Finally, we assess the current state of the field by reviewing established Performance Levels and Evaluation benchmarks and conclude by addressing the key Challenges and Future Directions that will shape the next generation of AI-driven scientific inquiry.

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

2 Capabilities and Core Functions of LLM Agents in Science

LLM agents in science operate across a sophisticated spectrum of capabilities that mirror the traditional scientific method. At their core, these systems can leverage vast interdisciplinary knowledge to break down information barriers and propose scientific hypotheses that have been validated against existing literature [24]. The agents demonstrate remarkable capacity to generate scientifically plausible and potentially novel hypotheses by combining their extensive domain knowledge with advanced reasoning capabilities [30] [34].

A key advancement is the agents' ability to integrate with external tools and scientific simulators, enabling automated statistical discovery and reasoning [30]. This integration allows LLM agents to move beyond theoretical hypothesis generation into practical experimentation and validation. For example, systems like FunSearch have demonstrated the ability to make genuine discoveries for established open problems by searching for programs that describe how to solve problems rather than what the solutions are [27].

The scientific research pipeline has been transformed as LLM agents can now collaborate across all critical stages including hypothesis generation, experimental design, data acquisition, and analysis [31]. These agents can interface with experimental data sources through programming execution, allowing for real-world experimentation and validation [25]. Domain-specific implementations like ChemCrow have shown how agents can autonomously plan and execute complex tasks such as chemical syntheses and guide the discovery of novel compounds [25] [5].

Recent research has established a three-level taxonomy for LLM involvement in scientific discovery: LLM as Tool for specific supervised tasks, LLM as Analyst for complex autonomous processing, and LLM as Scientist for fully autonomous research conduct from hypothesis formulation through result interpretation [46]. The most advanced capability involves autonomous knowledge generation, where agents synthesize data from multiple sources to propose novel insights, extrapolate trends, infer causality, and develop testable hypotheses, transforming them from passive information consumers into active contributors to scientific discovery [18].

> LLM agents can perform the full spectrum of scientific research tasks, from generating novel hypotheses and designing experiments to analyzing data and making discoveries. They function at three levels: as tools for specific tasks, as analysts for complex processing, or as autonomous scientists capable of conducting entire research workflows.

3 Multi-Agent Systems and Frameworks

The evolution toward multi-agent systems represents a significant advancement in autonomous scientific discovery, where specialized LLM agents collaborate to tackle complex research challenges that exceed the capabilities of individual agents. These frameworks harness what researchers describe as a "swarm of intelligence" similar to biological systems, enabling unprecedented scale, precision, and exploratory power that surpasses traditional human-driven research methods [12] [11].

Modern multi-agent scientific frameworks employ sophisticated role-based architectures where distinct agents assume specialized functions. The SciAgents framework exemplifies this approach by deploying agents with specific expertise as "Ontologist," "Scientist," and "Critic" to collectively generate and refine scientific hypotheses, orchestrating these ChatGPT-4-based agents around ontological knowledge graphs that encode relationships between scientific concepts [20]. Similarly, systems like CellAgent implement hierarchical decision-making mechanisms with planner, executor, and evaluator roles, incorporating self-iterative optimization to ensure output quality [8] [37].

Several notable frameworks have demonstrated end-to-end autonomous research capabilities. Agent Laboratory accepts human-provided research ideas and progresses through literature review, experimentation, and report writing stages, achieving an 84% reduction in research expenses compared to previous methods while enabling human feedback integration at each stage [29]. The Virtual Lab framework employs an LLM principal investigator guiding specialized agent teams with different scientific backgrounds, successfully designing functional nanobodies for SARS-CoV-2 variants through experimental validation [28] [32].

Advanced multi-agent systems are achieving remarkable discovery efficiency through sophisticated coordination mechanisms. The PiFlow framework treats scientific discovery as a structured uncertainty reduction problem, demonstrating a 73.55% increase in discovery efficiency and 94.06% enhancement in solution quality compared to single-agent systems across nanomaterials, bio-molecules, and superconductor research domains [23]. Other systems like IDVSCI incorporate Dynamic Knowledge Exchange mechanisms and Dual-Diversity Review paradigms to simulate heterogeneous expert evaluation, consistently outperforming existing frameworks in autonomous research tasks [42].

The integration of specialized tools and domain expertise enables these multi-agent systems to conduct sophisticated interdisciplinary research. Recent implementations have successfully generated thousands of structured hypotheses from vast literature databases, with rigorous evaluation processes identifying feasible, useful, and novel research directions [47] [40]. Contemporary frameworks like NovelSeek have achieved significant performance improvements across multiple scientific fields with dramatically reduced time costs, demonstrating accuracy increases from 27.6% to 35.4% in reaction yield prediction within just 12 hours [45].

> Multiple specialized AI agents work together in teams to conduct scientific research, with different agents handling specific roles like hypothesis generation, experiment design, and results evaluation. These collaborative frameworks achieve better research outcomes than single agents and can autonomously discover new materials, drugs, and scientific insights across diverse domains.

4 Domain-Specific Applications

4.1 Biology and Biomedicine

- *Protein Design*: ProtAgents enables collaborative design of novel proteins with targeted mechanical properties through dynamic multi-agent environments that combine knowledge retrieval, structure analysis, and physics-based simulations [10]

- *Single-Cell Analysis*: CellAgent automates scRNA-seq data processing with hierarchical decision-making mechanisms coordinating planner, executor, and evaluator roles, dramatically reducing workload for biological data analysis [37]

- *Genetic Research*: BioDiscoveryAgent autonomously designs genetic perturbation experiments, outperforming traditional methods in identifying genes linked to specific phenotypes and improving prediction accuracy [26]

- *Multi-Omics Analysis*: AutoBA leverages LLMs to automate bioinformatics analysis using established libraries to generate new biological insights [2]

4.2 Chemistry and Drug Discovery

- *Chemical Synthesis*: The notable ChemCrow system integrates 18 expert-designed tools with GPT-4 to autonomously plan and execute syntheses of insect repellents and organocatalysts while guiding discovery of novel chromophores [5]

- *Drug Development*: DrugAssist performs interactive molecule optimization through human-machine dialogue, achieving leading results in both single and multiple property optimization tasks [41]. DrugPilot demonstrates exceptional performance with task completion rates of 98.0%, 93.5%, and 64.0% for simple, multi-tool, and multi-turn drug discovery scenarios respectively [17]

- *Experimental Automation*: Coscientist combines LLMs to autonomously plan, design, and execute scientific experiments, successfully demonstrating catalyzed chemical reactions while addressing safety concerns [26]

4.3 Materials Science

- *Autonomous Synthesis*: The notable A-LAB system discovered and synthesized 41 novel compounds from 58 targets in 17 days of continuous operation, combining computations, literature data, and active learning for inorganic powder synthesis [33]

- *Alloy Design*: AtomAgents uses multi-agent frameworks combining physics-based simulations and multi-modal data integration for autonomous alloy discovery [13]

- *Crystal Structure Generation*: MatLLMSearch demonstrates that pre-trained LLMs can generate stable crystal structures without fine-tuning, achieving 78.38% metastable rate validated by machine learning potentials [9]

- *Data Extraction*: Eunomia autonomously extracts and structures experimental datasets from scientific literature, achieving performance comparable to state-of-the-art fine-tuned materials information extraction methods [1]

4.4 Healthcare and Clinical Applications

- *Speech-Language Pathology*: Specialized systems successfully identified 2,421 interventions from 64,177 research articles, creating publicly accessible intervention knowledge bases with significant community benefit [14]

- *Pharmaceutical Research*: AI co-scientist systems demonstrate empirically validated effectiveness in pharmaceutical repurposing, target discovery, and antimicrobial resistance research through multi-agent tournament-based evolutionary processes [26]

4.5 Cross-Domain Scientific Research

- *General Scientific Discovery*: The AI Scientist framework enables fully automated scientific discovery where LLMs independently generate ideas, execute experiments, write papers, and undergo review processes across multiple research fields [26]

- *Tool-Augmented Reasoning*: SciAgent systems retrieve, understand, and use specialized tools for scientific problem solving across five scientific domains, with SciAgent-Llama3-8B surpassing comparable LLMs by more than 8.0% in absolute accuracy [21]

> LLM agents are making real discoveries across many scientific fields, from finding new materials and drugs to analyzing biological data and designing proteins. These specialized systems have successfully identified thousands of research interventions, synthesized novel compounds, and automated complex experiments in chemistry, biology, materials science, and healthcare.

5 End-to-End Autonomous Research Systems

The development of end-to-end autonomous research systems represents the pinnacle of LLM-driven scientific discovery, where complete research workflows are automated from initial conception through final publication. Agent Laboratory exemplifies this capability by accepting human-provided research ideas and progressing through three comprehensive stagesliterature review, experimentation, and report writingto produce complete research outputs including code repositories and research reports while enabling user feedback at each stage [29]. This system achieves remarkable efficiency gains, demonstrating an 84% reduction in research expenses compared to previous autonomous research methods while generating machine learning code that achieves state-of-the-art performance [29].

Several pioneering frameworks have demonstrated successful end-to-end scientific discovery capabilities across diverse domains. The AI Scientist framework performs fully automated research in machine learning, including problem definition, experimental execution, code writing, and paper production with automated peer review [28] [45]. The enhanced AI Scientist-V2 incorporates agent tree search and vision-language model feedback, achieving the milestone of producing the first workshop paper fully generated and peer-reviewed by AI [45].

Real-world validation of these systems has produced tangible scientific breakthroughs. The Virtual Lab system employs an LLM principal investigator guiding specialized agent teams to design functional nanobody binders for SARS-CoV-2 variants, with experimental validation revealing promising binding profiles and two nanobodies showing improved binding to recent viral variants [28] [32]. Similarly, AI Co-Scientist has demonstrated empirically validated effectiveness in biomedical domains including drug repurposing and novel target identification through multi-agent systems employing "generate-debate-evolve" strategies [45].

The automation extends to physical experimentation through systems like ORGANA, which integrates decision-making and perception tools to automate diverse chemistry experiments while collaborating with chemists via LLMs to define objectives and generate detailed experiment logs [19] [7]. These robotic systems demonstrate over 50% reduction in user frustration and physical demand while saving researchers an average of 80.3% of their time [7].

Performance metrics across multiple scientific domains showcase the effectiveness of these autonomous systems. NovelSeek achieved significant accuracy improvements in just hours of processing: reaction yield prediction increased from 27.6% to 35.4% in 12 hours, enhancer activity prediction rose from 0.65 to 0.79 in 4 hours, and 2D semantic segmentation precision advanced from 78.8% to 81.0% in 30 hours [45]. These systems span the entire research pipeline from idea generation and experimental design to code implementation and academic paper drafting [39] [16] [44] [3] [15] [35].

> Complete autonomous research systems can now handle the entire scientific process from start to finish, taking in research ideas and producing full papers, code, and experimental results. These systems have successfully created functional discoveries like nanobodies and reduced research costs by up to 84% while maintaining scientific rigor.

6 Performance Levels and Evaluation

The evaluation of LLM agents in scientific discovery has evolved to include sophisticated performance taxonomies and comprehensive benchmarking frameworks that assess both current capabilities and future potential. A formal five-level performance hierarchy has been established, ranging from basic scientific tasks to paradigm-shifting discoveries [22]. At Level 3, agents demonstrate the ability to make novel scientific contributions worthy of publication at top conferences, while Level 4 encompasses groundbreaking contributions meriting oral presentations or best paper awards [22]. The highest Level 5 represents the ultimate goal: agents capable of pursuing long-term research agendas and producing paradigm-shifting breakthroughs worthy of Nobel or Turing prizes over extended periods [22].

Current evaluation frameworks reveal both the promise and limitations of existing systems. The Auto-Bench benchmark challenges LLMs to conduct human-like scientific research through causal graph discovery, requiring models to uncover hidden structures and make optimal decisions with valid justifications [6]. Testing state-of-the-art models including GPT-4, Gemini, Qwen, Claude, and Llama reveals significant performance degradation as problem complexity increases, highlighting important gaps between machine and human intelligence [6].

Real-world applications demonstrate impressive quantitative results in hypothesis generation and evaluation. Multi-agent systems have successfully processed massive datasets, with one implementation analyzing 66,000 scientific abstracts to produce 1,000 structured hypotheses [47]. Rigorous evaluation of these hypotheses revealed that 243 were deemed feasible based on current scientific knowledge, 175 demonstrated practical utility, and 12 stood out as highly novel contributions [47]. These systems employ sophisticated evaluation mechanisms including retrieval-augmented generation, tree-of-thoughts reasoning, and LLM-as-a-judge frameworks to ensure only the most promising hypotheses emerge from the discovery process [47] [40].

> Researchers have defined five performance levels for LLM scientific agents, from basic hypothesis generation to Nobel Prize-worthy breakthroughs, with current systems achieving notable success in mid-level tasks but showing significant performance drops as problem complexity increases. Evaluation frameworks now test agents on causal discovery, hypothesis generation, and multi-step reasoning across thousands of scientific problems.

7 Challenges and Future Directions

Despite the remarkable progress in LLM-driven scientific discovery, significant challenges remain that limit current systems' effectiveness and point toward critical areas for future development.

Comprehensive evaluation of state-of-the-art models including GPT-4, Gemini, Qwen, Claude, and Llama reveals a consistent pattern: performance drops significantly as problem complexity increases, highlighting an important gap between machine and human intelligence that future LLM development must address [6]. This performance degradation becomes particularly pronounced in tasks requiring causal graph discovery, where models must uncover hidden structures and make optimal decisions with valid justifications through iterative refinement processes [6].

A major frontier involves building more robust LLM agents that can effectively plan, reason, and interact with both humans and specialized scientific tools. The integration of LLMs into agent-based frameworks requires coordination with external tools such as retrosynthesis engines, docking software, and laboratory automation platforms to complete complex multi-step discovery workflows [36] [240288398 | García-Ortegón et al. | 2021 | Citations: 85]. These enhanced agent systems could potentially close the loop between computational prediction and experimental validation, enabling more flexible and goal-directed molecular design while accelerating the iterative discovery process [36] [240288398 | García-Ortegón et al. | 2021 | Citations: 85].

The development of causally-aware LLM agents represents another critical advancement area, with systems like MRAgent demonstrating the ability to autonomously scan literature, identify potential exposure-outcome pairs, execute causal inference analyses, and generate comprehensive reports [4] [38]. Future enhancements in AI-driven hypothesis generation will require agents to synthesize information from literature, structured databases, and experimental data to propose testable causal hypotheses, leveraging LLMs' strength in generating causal arguments based on their vast training data [4].

Advanced applications are emerging in target identification and validation, where causal agents integrate LLM-driven reasoning with data-driven causal discovery methods applied to omics data, identifying potential causal genes or pathways implicated in diseases with comprehensive explanations for their proposed roles [4] [43]. The integration of automated experiment analysis, including vision-based agents that can detect drug-cell interactions in microscopy images without task-specific training, promises to streamline experimental workflows and collectively shorten research cycles while prioritizing experiments based on causal plausibility [4].

> Current LLM agents face significant challenges as scientific problems become more complex, showing performance drops when dealing with intricate causal relationships and multi-step reasoning. Future development focuses on creating more robust agent frameworks that can better integrate computational predictions with experimental validation and handle complex causal discovery tasks.

Acknowledgements

Generative AI has been used to prepare this manuscript.

References

[1] Mehrad Ansari and S. M. Moosavi. Agent-based Learning of Materials Datasets from Scientific Literature. *arXiv.org*, 2023.

[2] Reza Averly, Frazier N. Baker, and Xia Ning. LIDDIA: Language-based Intelligent Drug Discovery Agent. *arXiv.org*, 2025.

[3] Jinheon Baek, S. Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models. *North American Chapter of the Association for Computational Linguistics*, 2024.

[4] Adib Bazgir, Amir Habibdoust Lafmajani, and Yuwen Zhang. Beyond Correlation: Towards Causal Large Language Model Agents in Biomedicine. *arXiv.org*, 2025.

[5] Andrés M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and P. Schwaller. Augmenting large language models with chemistry tools. *Nat. Mac. Intell.*, 2023.

[6] Tingting Chen, Srinivas Anumasa, Beibei Lin, Vedant Shah, Anirudh Goyal, and Dianbo Liu. Auto-Bench: An Automated Benchmark for Scientific Discovery in LLMs. *arXiv.org*, 2025.

[7] Kourosh Darvish, Marta Skreta, Yuchi Zhao, Naruki Yoshikawa, Sagnik Som, Miroslav Bogdanovic, Yang Cao, Han Hao, Haoping Xu, Alán Aspuru-Guzik, Animesh Garg, and F. Shkurti. ORGANA: A Robotic Assistant for Automated Chemistry Experimentation and Characterization. *Matter*, 2024.

[8] Shangheng Du, Jiabao Zhao, Jinxin Shi, Zhentao Xie, Xin Jiang, Yanhong Bai, and Liang He. A Survey on the Optimization of Large Language Model-based Agents. *arXiv.org*, 2025.

[9] Jingru Gan, Peichen Zhong, Yuanqi Du, Yanqiao Zhu, Chenru Duan, Haorui Wang, Carla P. Gomes, Kristin A. Persson, Daniel Schwalbe-Koda, and Wei Wang. Large Language Models Are Innate Crystal Structure Generators. *arXiv.org*, 2025.

[10] Alireza Ghafarollahi and Markus J. Buehler. ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 2024.

[11] Alireza Ghafarollahi and Markus J. Buehler. SciAgents: Automating Scientific Discovery Through Bioinspired MultiAgent Intelligent Graph Reasoning. *Advances in Materials*, 2024.

[12] Alireza Ghafarollahi and Markus J. Buehler. SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv.org*, 2024.

[13] Alireza Ghafarollahi and Markus J. Buehler. Automating alloy design and discovery with physics-aware multimodal multiagent AI. *Proceedings of the National Academy of Sciences of the United States of America*, 2025.

[14] Yuting Hu, Dancheng Liu, Qingyun Wang, Charles Yu, Heng Ji, and Jinjun Xiong. Automating Intervention Discovery from Scientific Literature: A Progressive Ontology Prompting and Dual-LLM Framework, 2024.

[15] Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. Autonomous LLM-driven research from data to human-verifiable research papers. *NEJM AI*, 2024.

[16] Mina Lee, Percy Liang, and Qian Yang. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. *International Conference on Human Factors in Computing Systems*, 2022.

[17] Kun Li, Zhennan Wu, Shoupeng Wang, and Wenbin Hu. DrugPilot: LLM-based Parameterized Reasoning Agent for Drug Discovery. *arXiv.org*, 2025.

[18] Chengwei Liu, Chong Wang, Jiayue Cao, Jingquan Ge, Kun Wang, Lvye Zhang, Ming-Ming Cheng, Penghai Zhao, Tianlin Li, Xiaojun Jia, Xiang Li, Xinfeng Li, Yang Liu, Yebo Feng, Yihao Huang, Yijia Xu, Yuqiang Sun, Zhe-Xu Zhou, and Zhengzi Xu. A Vision for Auto Research with LLM Agents. *arXiv.org*, 2025.

[19] Fan Liu, Zherui Yang, Cancheng Liu, Tianrui Song, Xiaofeng Gao, and Hao Liu. MM-Agent: LLM as Agents for Real-world Mathematical Modeling Problem. *arXiv.org*, 2025.

[20] Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, Rongcheng Tu, Xiaoming Luo, Wei Ju, Zhiping Xiao, Yifan Wang, Mengxue Xiao, Chenwu Liu, Jingyang Yuan, Shichang Zhang, Yiqiao Jin, Fan Zhang, Xianhong Wu, Hanqing Zhao, Dacheng Tao, Philip S. Yu, and Ming Zhang. Large Language Model Agent: A Survey on Methodology, Applications and Challenges. *arXiv.org*, 2025.

[21] Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, and Aixin Sun. SciAgent: Tool-augmented Language Models for Scientific Reasoning. *Conference on Empirical Methods in Natural Language Processing*, 2024.

[22] Deepak Nathani, Lovish Madaan, Nicholas Roberts, Niko lay Bashlykov, A. Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Foerster, Yoram Bachrach, William Yang Wang, and R. Raileanu. MLGym: A New Framework and Benchmark for Advancing AI Research Agents. *arXiv.org*, 2025.

[23] Yingming Pu, Tao Lin, and Hongyu Chen. PiFlow: Principle-aware Scientific Discovery with Multi-Agent Collaboration. *arXiv.org*, 2025.

[24] Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhangren Chen, and Bowen Zhou. Large Language Models are Zero Shot Hypothesis Proposers. *arXiv.org*, 2023.

[25] Chandan K. Reddy and Parshin Shojaee. Towards Scientific Discovery with Generative AI: Progress, Opportunities, and Challenges. *AAAI Conference on Artificial Intelligence*, 2024.

[26] Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. Towards Scientific Intelligence: A Survey of LLM-based Scientific Agents. *arXiv.org*, 2025.

[27] Bernardino Romera-Paredes, M. Barekatain, Alexander Novikov, Matej Balog, M. P. Kumar, Emilien Dupont, Francisco J. R. Ruiz, J. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, Alhussein Fawzi, Josh Grochow, Andrea Lodi, Jean-Baptiste Mouret, Talia Ringer, and Tao Yu. Mathematical discoveries from program search with large language models. *The Naturalist*, 2023.

[28] Samuel Schmidgall and Michael Moor. AgentRxiv: Towards Collaborative Autonomous Research. *arXiv.org*, 2025.

[29] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent Laboratory: Using LLM Agents as Research Assistants. *arXiv.org*, 2025.

[30] Parshin Shojaee, Kazem Meidani, Shashank Gupta, A. Farimani, and Chandan K. Reddy. LLM-SR: Scientific Equation Discovery via Programming with Large Language Models. *International Conference on Learning Representations*, 2024.

[31] Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[32] Kyle Swanson, Wesley Wu, Nash L. Bulaong, J. Pak, and James Y. Zou. The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation. *bioRxiv*, 2024.

[33] N. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, T. He, David Milsted, Matthew J. McDermott, Max C. Gallant, E. D. Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin A. Persson, Yan Zeng, and G. Ceder. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 2023.

[34] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific Inspiration Machines Optimized for Novelty. *Annual Meeting of the Association for Computational Linguistics*, 2023.

[35] Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. AutoSurvey: Large Language Models Can Automatically Write Surveys. *Neural Information Processing Systems*, 2024.

[36] Ziqing Wang, Kexin Zhang, Zihan Zhao, Yibo Wen, Abhishek Pandey, Han Liu, and Kaize Ding. A Survey of Large Language Models for Text-Guided Molecular Discovery: from Molecule Generation to Optimization. *arXiv.org*, 2025.

[37] Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, Shaoqing Jiao, and Jiajie Peng. CellAgent: An LLM-driven Multi-Agent Framework for Automated Single-cell Data Analysis. *arXiv.org*, 2024.

[38] Wei Xu, Gang Luo, Weiyu Meng, Xiaobing Zhai, Ke Zheng, Ji Wu, Yanrong Li, Abao Xing, Junrong Li, Zhifan Li, Ke Zheng, and Kefeng Li. MRAgent: an LLM-based automated agent for causal knowledge discovery in disease via Mendelian randomization. *Briefings Bioinform.*, 2025.

[39] Shuo Yan, Ruochen Li, Ziming Luo, Zimu Wang, Daoyang Li, Liqiang Jing, Kaiyu He, Peilin Wu, George Michalopoulos, Yue Zhang, Ziyang Zhang, Mian Zhang, Zhiyu Chen, and Xinya Du. LMR-BENCH: Evaluating LLM Agent's Ability on Reproducing Language Modeling Research. *arXiv.org*, 2025.

[40] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, T. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Neural Information Processing Systems*, 2023.

[41] Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xian Zeng. DrugAssist: a large language model for molecule optimization. *Briefings Bioinform.*, 2023.

[42] Weilun Yu, Shixiang Tang, Yonggui Huang, Nanqing Dong, Li Fan, Honggang Qi, Wei Liu, Xiaoli Diao, Xi Chen, and Wanli Ouyang. Dynamic Knowledge Exchange and Dual-diversity Review: Concisely Unleashing the Potential of a Multi-Agent Research Team. *arXiv.org*, 2025.

[43] Haolong Zeng, Chaoyi Yin, Chunyang Chai, Yuezhu Wang, Qi Dai, and Huiyan Sun. Cancer gene identification through integrating causal prompting large language model with omics datadriven causal inference. *Briefings Bioinform.*, 2025.

[44] Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. CodeAgent: Enhancing Code Generation with Tool-Integrated Agent Systems for Real-World Repo-level Coding Challenges. *Annual Meeting of the Association for Computational Linguistics*, 2024.

[45] NovelSeek Team Bo Zhang, Shi Feng, Xiangchao Yan, Jiakang Yuan, Zhiyin Yu, Xiaohan He, Songtao Huang, Shaowei Hou, Zheng Nie, Zhilong Wang, Jinyao Liu, Runmin Ma, Tianshuo Peng, Peng Ye, Dongzhan Zhou, Shufei Zhang, Xiaosong Wang, Yilan Zhang, Meng Li, Zhongying Tu, Xiangyu Yue, Wangli Ouyang, Bowen Zhou, and Lei Bai. NovelSeek: When Agent Becomes the Scientist - Building Closed-Loop System from Hypothesis to Verification. *arXiv.org*, 2025.

[46] Tianshi ZHENG, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From Automation to Autonomy: A Survey on Large Language Models in Scientific Discovery. *arXiv.org*, 2025.

[47] Yoel Zimmermann, Adib Bazgir, Alexander Al-Feghali, Mehrad Ansari, L. C. Brinson, Chiang Yuan, Defne Çirci, Min-Hsueh Chiu, Nathan Daelman, Matthew L Evans, Abhijeet Gangan, Janine George, Hassan Harb, Ghazal Khalighinejad, S. Khan, Sascha Klawohn, Magdalena Lederbauer, Soroush Mahjoubi, Bernadette Mohr, S. M. Moosavi, A. Naik, Aleyna Beste Ozhan, D. Plessers, Aritra Roy, Fabian Schoppach, Philipp Schwaller, Carla Terboven, Katharina Ueltzen, Shang Zhu, Jan Janssen, Calvin Li, Ian T. Foster, and B. Blaiszik. 34 Examples of LLM Applications in Materials Science and Chemistry: Towards Automation, Assistants, Agents, and Accelerated Scientific Discovery. *arXiv.org*, 2025.