

Preview

Agentifying the scientific method

Jiheon Kim^{1,2,3} and Edward H. Sargent^{1,3,4,5,*}¹The Edward S. Rogers Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada²Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada³The Alliance for AI-Accelerated Materials Discovery (A3MD), Toronto, ON, Canada⁴Department of Chemistry, Northwestern University, Evanston, IL, USA⁵Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA*Correspondence: ted.sargent@northwestern.edu<https://doi.org/10.1016/j.matt.2026.103045>

Automation in materials research has targeted the making; the scientific method—hypothesizing, designing discriminating experiments, judging what the data mean—has remained human. This preview highlights a study by Ghareeb, Chang, and colleagues in which a multi-agent system takes on those stages and asks what that means for materials discovery.

Automation in materials research has mostly focused on the physical loop. Robotic sample handling, parallelized workflows, and in-line characterization have compressed the time required to prepare and measure a sample and have removed much of the manual labor involved.^{1,2} Yet the rate-limiting step in discovery is often not the making. It is the scientific method itself—above all, framing a hypothesis worth testing and designing the experiment that will actually discriminate between competing explanations, bracketed by the reading of a scattered literature beforehand and the extraction of meaning from noisy data after. This is work that has remained almost entirely human.

Writing in *Nature*, Ghareeb, Chang, and colleagues introduce Robin, a multi-agent system that automates the intellectual stages of a discovery cycle while leaving bench execution to human scientists.³ Given only the name of a disease, Robin surveyed the literature, selected a therapeutic strategy and an assay with which to test it, and nominated the compounds to screen. Human scientists then performed the experiments.

Three specialized agents perform the retrieval and the analysis, and Robin is the structured workflow that coordinates them (Figure 1A and 1B). Two of the agents, Crow and Falcon, are built on PaperQA2,⁴ and they search the literature and ground their answers in the retrieved text rather than in the model's parameters, so their citations stay attached to real papers. Crow returns concise re-

ports, and Falcon writes deeper evaluations that include the case against each candidate. The third agent, Finch, analyzes experimental data by writing and executing its own code. Because the same agent given the same data does not always reach the same conclusion, Robin launches eight independent Finch analyses and reconciles them into a consensus. Robin itself supplies the connective reasoning. It composes the questions, ranks the returned hypotheses in a large language model-judged pairwise tournament, and determines what each result implies for the round that follows.

The authors deployed Robin against dry age-related macular degeneration, the leading cause of irreversible sight loss in developed countries. Robin reviewed 151 papers to shortlist ten candidate disease mechanisms, ranked them, and selected the strategy of restoring the debris-clearing capacity of the retinal cell layer that fails in this disease. It specified the assay that would measure this capacity cell by cell, then reviewed some 400 further papers to nominate and rank thirty existing drugs. The literature-agent calls behind all of this cost under \$11. The authors estimate that the equivalent intellectual work (searching, synthesizing, ideating, and designing) would occupy a human scientist for 359 to 424 h.

Then human scientists re-entered the loop. They advanced the top five candidates from the ranking and wrote their own protocol around the assay Robin had specified, adapting several of its choices to what the laboratory had on

hand. They then performed the experiment and returned the raw data files to Robin with a brief instruction naming the analysis to be run. Finch carried out the remainder, writing and executing its own analysis code, and a human reanalysis of the same data confirmed its conclusion. One compound stood out: Y-27632, a research-grade inhibitor of Rho kinase (Figure 1C; Robin round 1). Robin then proposed a follow-up experiment of its own, profiling gene expression in treated cells to ask how the hit acts. The result was consistent with the compound's known mode of action.

The second round is where the loop closes. Robin incorporated these results into its next literature search and, reasoning now from experimental evidence, proposed a new set of candidates. Its leading proposal, ripasudil, kept the validated target but changed the molecule. As a drug already approved in Japan for another disease of the same organ, it carries a safety record that a research compound lacks. Of ten candidates tested, ripasudil outperformed Y-27632, a gain confirmed by a dose-response experiment. A final re-screen in primary human cells from an elderly donor, under conditions much closer to the diseased tissue, upheld this result and added a second hit, KL001, which achieves the same effect by a route the authors believe to be untried (Figure 1C; Robin round 2). Two rounds therefore traced a path from a plausible mechanism to a working research compound to an approved drug repurposed for a new disease, with



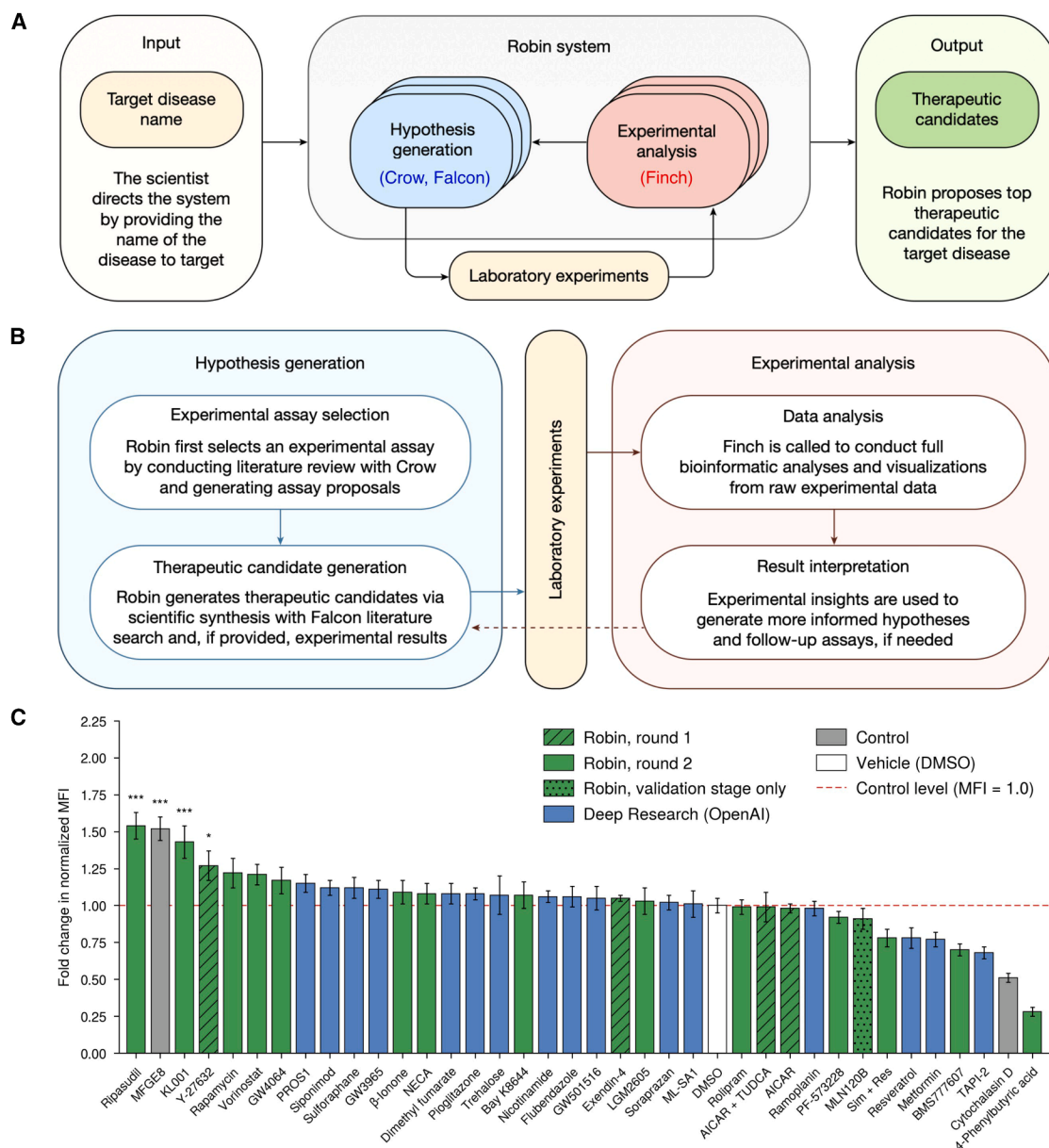


Figure 1. Robin, a multi-agent system that automates the reasoning stages of a discovery cycle

(A) System overview. A scientist names a target disease, and Robin alternates between hypothesis generation (Crow, Falcon) and experimental analysis (Finch), with laboratory experiments performed by human scientists in between.

(B) One round in detail. Robin selects an assay, generates therapeutic candidates from literature searches and any preceding experimental results, and feeds Finch's analysis into the next round (dashed arrow).

(C) Compounds from both Robin rounds screened against those of a general-purpose research agent, in primary human retinal cells from an aged donor. Bars show fold change in mean fluorescence intensity relative to vehicle; error bars show the SEM ($n = 4$). One-sided Dunnett test: $*p < 0.05$, $***p < 0.001$.

(A) and (B) adapted and (C) replotted from data in Ghareeb et al.³ with permission from Springer Nature.

the machine performing the reading and the proposing at every step and the humans applying domain judgment and carrying out the experiments.

The authors also validate the architecture of Robin itself. Replacing the literature-search agent with OpenAI's o4-mini answering directly raised the proportion of

fabricated references from none to nearly half. The retrieval agents do not make the system a better reasoner. They make its output verifiable, keeping each claim connected to a real paper. Finch was evaluated as well. Scored against a rubric written by human experts, its analyses met nearly every criterion. The most direct compari-

son came last. Deep Research by OpenAI, a leading general-purpose research agent, was given the same task and asked for the same number of candidates, and none of the drugs it proposed proved active (Figure 1C; Deep Research).

The authors read these results as evidence of a particular kind of capability.

Robin's hypotheses are novel not because any single insight is new but because the system connects insights that already exist, scattered across the literature or belonging to separate fields. They are also clear about what remains undone. Robin writes experimental outlines rather than executable protocols, so a human expert must still translate each proposal into a procedure the laboratory can run, and Finch depends on analysis instructions written by domain experts. The authors aim to address both as a next step. Future versions are intended to produce protocols detailed enough to run with minimal human translation, and an analysis agent that writes its own instructions for a given type of data. These capabilities are also central to the autonomy that self-driving laboratories (SDLs) ultimately pursue.

This paradigm of reliable, literature-grounded synthesis, they argue, is not specific to drug repurposing, and the authors name "materials science" among the fields to which it could extend. We agree with this expectation, and the extension will be a genuine test, because the two discovery problems are structurally distinct. Drug repurposing is, by design, a selection problem. The candidates already exist as discrete, nameable molecules, and the question is which of them to try. Discovering a heterogeneous catalyst or a photovoltaic absorber is a generation problem.⁵ Candidate materials must first be synthesized, and composition, structure, and processing together define both what is made and how it performs in the final device. None of this argues against agentic discovery in materials science. It argues for adopting it as Robin actually operated, with humans in the loop, rather than as a fully autonomous pipeline.

Full autonomy remains the ambition of SDL researchers across many fields,⁶

but for the near term, and possibly much longer, the practical path will be a deliberate division of labor in which AI agents read and propose across volumes of literature that no specialist can hold; robots make and measure; and human judgment refines the agents' proposals into executable procedures, steers with domain expertise,^{7,8} and audits interpretation of data that are inevitably noisy. Robin's deepest lesson is that literature grounding and careful orchestration, not raw model power, are what made its loop work. Likewise for materials science, the task now extends beyond automating its far more complex experiments: agentic hypothesizing amid data that are noisy, nonlinearly coupled, and imperfect remains in its infancy, and bringing it to maturity will require agents, robots, and human experts working in concert.

ACKNOWLEDGMENTS

This work was supported by the Alliance for AI-Accelerated Materials Discovery (A3MD).

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE MANUSCRIPT PREPARATION PROCESS

The first draft of this preview, including its analysis and its argument, was conceived and written in full by J.K.; E.H.S. reviewed that draft and provided comments. During revision, the authors used Anthropic's Claude to help polish the manuscript: it proposed the title, suggested a reordering of the opening paragraph, repaired a malformed sentence in the concluding paragraph, and made copy edits for grammar and consistency. No AI-generated text entered the manuscript without author review. The authors reviewed and edited all such content and take full responsibility for the content of this publication.

REFERENCES

1. Burger, B., Maffettone, P.M., Gusev, V.V., Aitchison, C.M., Bai, Y., Wang, X., Li, X., Alston, B.M., Li, B., Clowes, R., et al. (2020). A mobile robotic chemist. *Nature* 583, 237–241. <https://doi.org/10.1038/s41586-020-2442-2>.
2. Dai, T., Vijayakrishnan, S., Szczepiński, F.T., Ayme, J.F., Simaei, E., Fellowes, T., Clowes, R., Kotopantov, L., Shields, C.E., Zhou, Z., et al. (2024). Autonomous mobile robots for exploratory synthetic chemistry. *Nature* 635, 890–897. <https://doi.org/10.1038/s41586-024-08173-7>.
3. Ghareeb, A.E., Chang, B., Mitchener, L., Yiu, A., Szostkiewicz, C.J., Shved, D., Gyimesi, G.J., Laurent, J.M., Wright, S.M., Razzak, M.T., et al. (2026). A multi-agent system for automating scientific discovery. *Nature* 655, 497–505. <https://doi.org/10.1038/s41586-026-10652-y>.
4. Skarlinski, M.D., Cox, S., Laurent, J.M., Braza, J.D., Hinks, M., Hammerling, M.J., Ponnampati, M., Rodrigues, S.G., and White, A.D. Language agents achieve superhuman synthesis of scientific knowledge. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2409.13740>.
5. Bati, A.S.R., Liu, C., Gilley, I.W., Musgrave, C.B., III, Maxwell, A., Steele, J.A., Yang, Y., Chen, H., Wan, H., Xu, J., et al. (2025). A chemically bonded monolayer interface enables enhanced thermal stability and efficiency in Pb-Sn perovskite solar cells. *Joule* 9, 102047. <https://doi.org/10.1016/j.joule.2025.102047>.
6. Szymanski, N.J., Rendy, B., Fei, Y., Kumar, R.E., He, T., Milsted, D., McDermott, M.J., Gallant, M., Cubuk, E.D., Merchant, A., et al. (2023). An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature* 624, 86–91. <https://doi.org/10.1038/s41586-023-06734-w>.
7. Kim, J., Mahesh, S., Lee, H.S., Dorakhan, R., Bai, Y., Imran, M., Li, K., Liu, Y., Kim, D., Park, S., et al. (2025). Accelerated discovery of CO₂-to-C₃-hydrocarbon electrocatalysts with human-in-the-loop. *Joule* 9, 102213. <https://doi.org/10.1016/j.joule.2025.102213>.
8. Adams, F., McDannald, A., Takeuchi, I., and Kusne, A.G. (2024). Human-in-the-loop for Bayesian autonomous materials phase mapping. *Matter* 7, 697–709. <https://doi.org/10.1016/j.matt.2024.01.005>.