

Accelerating scientific discovery with Co-Scientist

<https://doi.org/10.1038/s41586-026-10644-y>

Received: 20 March 2025

Accepted: 11 May 2026

Published online: 19 May 2026

Open access

 Check for updates

Juraj Gottweis^{1,8}, Wei-Hung Weng^{2,8}, Alexander Daryin^{1,8}, Tao Tu^{2,8}, Petar Sirkovic^{1,8}, Artiom Myaskovsky^{1,8}, Grzegorz Glowaty^{1,8}, Felix Weissenberger^{1,8}, Alessio Orlandi^{1,8}, Dan Popovici³, Anil Palepu³, Keran Rong², Ryutaro Tanno², Khaled Saab², Fan Zhang³, Jacob Blum⁴, Andrew Carroll³, Kavita Kulkarni³, Nenad Tomašev², Dina Zverinski¹, Ivor Rendulic¹, Elahe Vedadi², Florian Hasler¹, Luka Rimanic¹, Marina Boia¹, Ivan Budiselic¹, Ben Feinstein³, Mathias Bellaïche³, Tom Sheffer³, Jan Freyberg², Jeremy Ratcliff², Ottavia Bertolli², Katherine Chou³, Avinatan Hassidim³, Burak Gokturk¹, Amin Vahdat¹, Yuan Guan⁴, Vikram Dhillon⁵, Eeshit Dhaval Vaishnav⁶, Byron Lee⁶, Tiago R. D. Costa⁷, José R. Penadés⁷, Gary Peltz⁴, Yossi Matias³, James Manyika³, Demis Hassabis², Yunhan Xu², Pushmeet Kohli², Annalisa Pawlosky³, Alan Karthikesalingam² & Vivek Natarajan²

Scientific discovery is driven by scientists generating hypotheses for complex problems that undergo rigorous experimental validation. To augment this process, we introduce Co-Scientist, a multi-agent artificial intelligence (AI) system built on Gemini for structured scientific thinking and hypothesis generation. Co-Scientist aims to help scientists discover new original knowledge. Conditioned on their research objectives and previous scientific evidence, it formulates demonstrably novel research hypotheses for experimental verification. The system's design involves agents continuously generating, critiquing and refining hypotheses accelerated by scaling test-time compute. Key contributions include (1) a multi-agent architecture with an asynchronous task execution framework for flexible compute scaling, and (2) a tournament evolution process for self-improving hypotheses generation. Automated evaluations show continued benefits of test-time compute scaling, improving hypothesis quality over time. Although this is a general-purpose system, we focus the validation in three biomedical applications: drug repurposing; novel-target discovery¹; and explaining mechanisms of antimicrobial resistance². Specifically, Co-Scientist helped to identify new drug-repurposing candidates and synergistic combination therapies for acute myeloid leukaemia that were validated through *in vitro* experiments. These real-world validations demonstrate the potential of Co-Scientist to accelerate scientific discovery and usher in an era of AI-empowered scientists.

Researchers are faced with a breadth and depth conundrum. The complexity of scientific topics requires increasingly deep and specific subject matter expertise, while leaps in insight may still arise from broad knowledge bridging across disciplines^{3–5}. With the rapid rise in scientific publications and the development of numerous specialized technologies, mastery of both discipline-specific depth and trans-disciplinary insights can be challenging.

At the same time, there has been rapid technological progress in AI towards generally intelligent and collaborative systems, which might empower scientists in creatively traversing and expertly reasoning across disciplinary domains. Such systems are capable of advanced reasoning^{6–8}, multimodal understanding⁸ and agentic actions⁹, such as the ability to use tools to solve complex tasks over long time horizons. Furthermore, the trends with distillation¹⁰ and inference time compute costs^{8,11} indicate that such intelligent and general AI systems are rapidly

becoming more accessible. Motivated by the aforementioned unmet needs in the modern discovery process in science and medicine and building on the advancements in frontier AI¹², we develop and introduce Co-Scientist.

Co-Scientist is a structured scientific thinking engine designed to act as a collaborator to scientists and help accelerate the scientific discovery process. The system is a compound, multi-agent AI system¹³ building on Google's large language model (LLM) Gemini¹⁴, mirroring the reasoning process underpinning the scientific method¹⁵. Given a research goal specified in natural language, the system can search, learn and reason over relevant literature to synthesize previous work and propose novel, original research hypotheses and experimental protocols (Fig. 1a). Co-Scientist provides grounding for its recommendations by citing relevant literature, applying sound scientific reasoning and verifying its conclusions through external tools when applicable.

¹Google Cloud AI Research, Zurich, Switzerland. ²Google DeepMind, Mountain View, CA, USA. ³Google Research, Mountain View, CA, USA. ⁴Stanford University School of Medicine, Palo Alto, CA, USA. ⁵Houston Methodist Hospital, Houston, TX, USA. ⁶Sequome, San Francisco, CA, USA. ⁷Fleming Initiative and Imperial College London, London, UK. ⁸These authors contributed equally: Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi. [✉]e-mail: juro@google.com; ckbjimmy@google.com; pushmeet@google.com; apawlosky@google.com; alankarthi@google.com; natviv@google.com

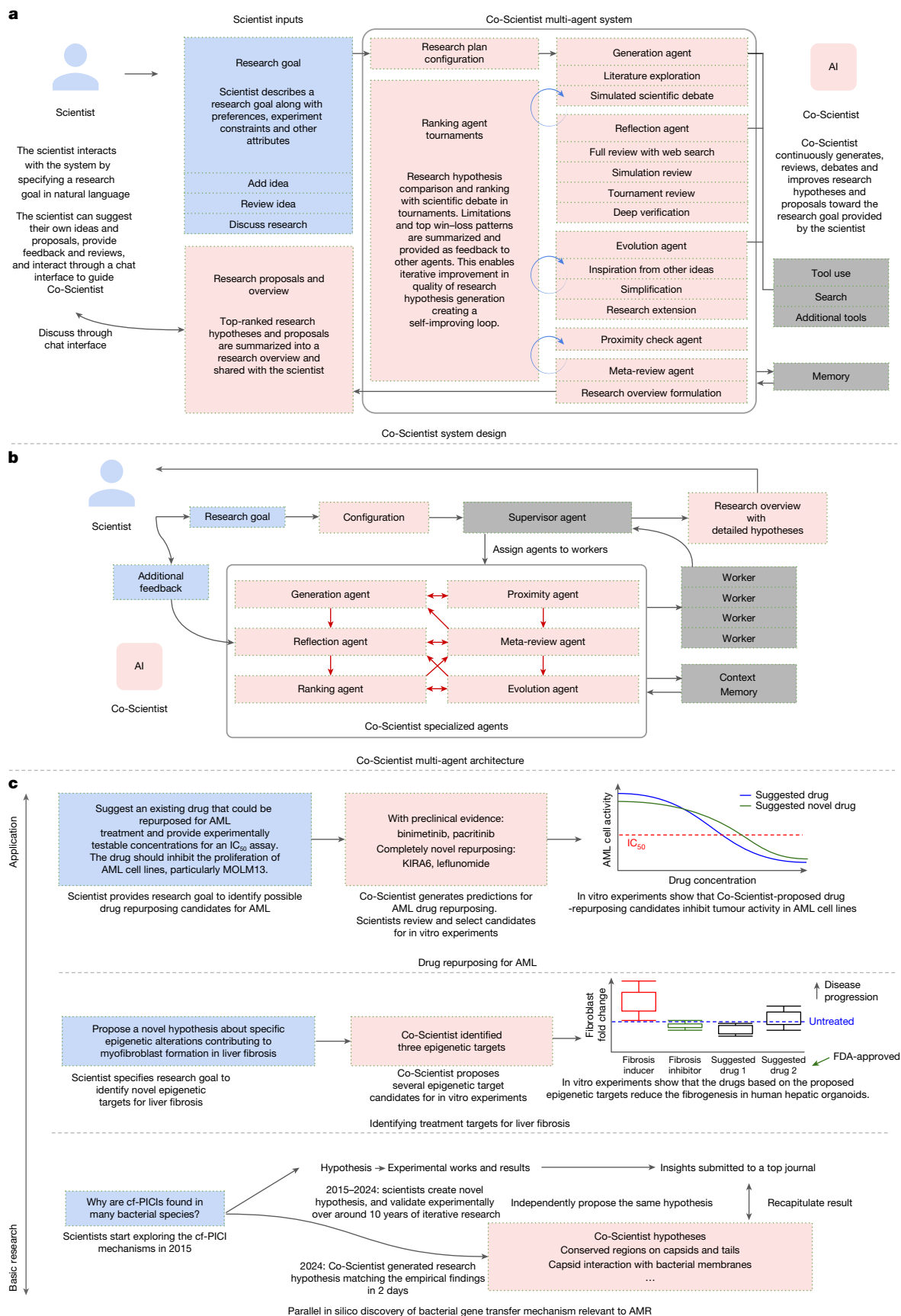


Fig. 1 | See next page for caption.

Fig. 1 | Summary of Co-Scientist design, its multi-agent architecture and experimental validation. **a**, Overview of the different components of Co-Scientist's structured scientific thinking engine—the multi-agent system, and its interaction paradigm with scientists. Given a research goal in natural language, Co-Scientist generates novel research hypotheses. The system uses specialized Gemini-based agents, including Generation, Reflection, Ranking, Evolution, Proximity (which evaluates relatedness) and Meta-review (which provides high level analysis) agents, to continuously generate, debate and evolve research hypotheses within a tournament framework. Feedback from the tournament enables iterative improvement, creating a self-improving loop towards novel and high-quality hypotheses for solving complex scientific problems. Co-Scientist uses tools, including web search and specialized AI models, to improve the grounding and quality of generated research hypotheses. Scientists can converse with Co-Scientist in natural language to specify research goals, incorporate constraints, provide feedback, steer and suggest new directions for explorations through the designated user interface.

Co-Scientist is purpose-built for a 'scientist-in-the-loop' collaborative paradigm. Scientists can specify their research goals in simple natural language and inform the system of desirable attributes and constraints for the proposed solutions. They can also actively interact with and steer the system, including directly suggesting initial ideas and hypotheses for the exploration, refining generated ideas or providing feedback through natural language chat.

Co-Scientist works through a substantial scaling of the test-time compute paradigm^{16–18}, implementing structured scientific thinking in a multi-agent setup to iteratively reason, evolve and improve the outputs as it gathers more knowledge (Fig. 1b). Underpinning the system are thinking and reasoning steps—notably a self-play-based scientific debate step for generating novel research hypotheses; tournaments that compare and rank hypotheses through the process of finding win and loss patterns; and an evolution process to improve their quality. Finally, the agentic nature of the system enables it to recursively self-critique its output and use tools such as web search and specialized AI models to provide itself with feedback to refine its hypotheses and research proposals.

Although Co-Scientist is general purpose and applicable across scientific disciplines, here we validate it in three impactful areas of biomedicine with varied complexity: drug repurposing for cancer; novel treatment target discovery for liver fibrosis; and identification of mechanistic explanations for antimicrobial resistance (AMR) (Fig. 1c).

Drug development remains an expensive and protracted process, with most new approvals requiring *de novo* discovery for each indication¹⁹. Systematic identification of new therapeutic indications for approved agents through drug repurposing offers a pragmatic strategy to accelerate development timelines and reduce attrition²⁰. Using Co-Scientist, we generated large-scale repurposing predictions validated through expert curation and *in vitro* assays. The system proposed several single-agent and combination therapies for AML that demonstrate selective cytotoxicity at clinically relevant concentrations. Beyond repurposing, Co-Scientist enables hypothesis generation for *de novo* target discovery, a process that is traditionally limited by the scale and uncertainty of biological inference. We applied Co-Scientist to liver fibrosis, for which it proposed and ranked novel epigenetic targets demonstrating significant anti-fibrotic activity and hepatocyte regeneration in human hepatic organoids¹. Finally, we examined bacterial gene transfer mechanisms related to AMR, a system-level challenge involving molecular mechanisms and evolutionary pressures². Researchers instructed the system to explore a topic their group had independently discovered, but not yet published. Co-Scientist was asked to hypothesize how capsid-forming phage-inducible chromosomal islands (cf-PICIs) exist across bacterial species. It independently proposed that cf-PICIs interact with diverse phage tails to expand host range, mirroring the researchers' unpublished experimental findings detailed in co-timed reports^{2,21}.

b, The underlying multi-agent architecture: a Supervisor agent parses the user's natural language research goal and dynamically allocates resources to specialized worker agents within an asynchronous task queue. The red boxes in the 'Co-Scientist specialized agents' part denote individual agents each with their own unique logic and role. The blue boxes indicate the scientist-in-the-loop inputs and feedback. The dark grey arrows represent the information flow through the Co-Scientist, and the red arrows represent the information feedback loop between the specialized agents. **c**, End-to-end validation of Co-Scientist across three biomedical problems of varying complexity: Co-Scientist proposed drug-repurposing candidates for AML (top), identified novel epigenetic targets for liver fibrosis (middle) and independently recapitulated a, then unpublished, co-timed discovery of a novel bacterial gene transfer mechanism relevant to AMR (bottom). All hypotheses generated by Co-Scientist were subsequently verified through independent *in vitro* laboratory experiments.

Key contributions

Introducing Co-Scientist

We develop and introduce Co-Scientist, a structured scientific thinking engine, that goes beyond literature summarization and deep research tools to assist scientists in generating knowledge, novel hypothesis generation, identifying unexpected connections and experimental planning.

Substantial scaling of the test-time compute paradigm for scientific reasoning

Co-Scientist is built on a Gemini-based multi-agent architecture, using an asynchronous task-execution framework. This framework enables the system to flexibly allocate computational resources to scientific reasoning, mirroring key aspects of the scientific method. Specifically, the system uses self-play strategies, including a scientific debate and a tournament-based evolution process, to iteratively refine hypotheses and research proposals creating a self-improving loop. Using automated evaluations across 15 complex expert curated open scientific goals, we demonstrate the benefits of scaling the test-time compute paradigm with Co-Scientist outperforming other state-of-the-art agentic and reasoning models in generating high-quality hypotheses for complex problems.

Expert-in-the-loop scientific workflow

Our system is designed for collaboration with scientists. The system can flexibly incorporate conversational feedback in natural language from scientists and co-develop, evolve and refine outputs.

End-to-end validation of Co-Scientist in important topics in biomedicine

We present end-to-end validation of AI-generated hypotheses through new empirical findings in three distinct and increasingly complex areas of biomedicine: drug repurposing, novel-target discovery and AMR (Table 1 and Supplementary Note 1).

Co-Scientist overview

Given a research goal, Co-Scientist generates hypotheses constrained by default criteria that include plausibility, novelty, testability and safety. At a high level, it uses an asynchronous multi-agent architecture whereby a team of agents co-operate to solve scientific problems and develop new hypotheses (Fig. 1b). It comprises a natural language interface for expert supervision, a task-execution framework for resource allocation, a suite of specialized agents (Generation, Reflection, Ranking, Evolution, Proximity and Meta-review) mirroring the scientific method and a persistent context memory for long-horizon reasoning. Detailed system configurations and agent mechanisms are fully described in the Methods.

Table 1 | Real-world applications in biomedicine for end-to-end validation of Co-Scientist

Application	Drug repurposing	Novel treatment target discovery	Explain mechanism of gene transfer evolution
Challenge	Complex search	Identifying novel targets	Understanding complex systems
Complexity	Medium	High	Very high
Scale	Moderate, data limited	Moderate, experiment limited	Large, data and computation limited
Unknown elements	Constrained	Large	Vast and dynamic

Summary of three scientific tasks selected to evaluate the hypothesis generation abilities of Co-Scientist. The chosen applications span varying biological disciplines and are categorized by four axes, inherent challenge (the primary scientific objective), complexity (the depth of reasoning required), scale (data availability and experimental feasibility) and unknown elements (the boundaries of the hypothesis search space). These progressively demanding tasks illustrate Co-Scientist’s generalizability and its ability to navigate both constrained problem spaces and open-ended explorations.

System analysis and evaluation

We first conducted the initial system evaluations to benchmark and verify the choice of the architecture and metrics underpinning Co-Scientist (Supplementary Note 2 and Supplementary Fig. 1). We performed an ablation study to investigate the contribution of each agentic component in Co-Scientist. We then analysed the impact of scaling test-time compute, and undertook a small-scale evaluation with domain experts to assess the quality of the system outputs. Finally, to assess the practical use of the system’s predictions, we performed end-to-end wet-laboratory validations (laboratory experiments) of Co-Scientist-generated hypotheses and research proposals in three key biomedical applications: drug repurposing, identifying new treatment targets and elucidating the mechanisms underlying AMR (Table 1 and Supplementary Note 1). The varying complexity and nature of these applications enable a more comprehensive assessment of the system. Notably, all three validations included the expert-in-the-loop concept.

Agent ablation analysis

Our ablation analyses (described in the Methods, Supplementary Note 3 and Supplementary Figs. 2–6) confirmed the importance of our multi-agent architecture and specialized prompting strategies for robust scientific reasoning. For example, granting the Reflection agent access to external search tools effectively prevented the hallucination of seemingly novel but implausible hypotheses, while using a scientific debate prompt in the Ranking agent substantially improved the ranking of hypotheses and reduced positional bias. Furthermore, iterative refinement by the Evolution agent substantially boosted hypotheses quality.

Scaling test-time compute improves scientific reasoning

To evaluate the effects of test-time compute scaling and Co-Scientist’s progress during iterative scientific reasoning and hypothesis generation, we measured the Elo ratings of Co-Scientist-generated hypotheses and proposals over the course of its thinking and computation (that is, the tournament of hypotheses). This analysis was done across 203 distinct research goals curated across broad scientific topics (predominantly in biomedicine, but also included other topics, such as mathematics and physics) and entered into Co-Scientist until 3 February 2025.

Co-Scientist’s research hypotheses and proposals were partitioned into ten temporal buckets of equal size. Each bucket corresponded to a sequential 10% of the total generation time, with the first bucket containing the earliest 10% of generated Co-Scientist results, while the tenth bucket comprised the most recent 10%. For each bucket, we determined the average Elo rating of the top-ten hypotheses and the maximum individual Elo rating (the best Elo). These average and best Elo ratings were averaged across 203 research goals and their corresponding tournaments. The resulting performance trends as seen in Fig. 2a, across both metrics, serve as a measure of Co-Scientist’s quality improvement as it spends more time in thinking and computation—the most recent hypotheses demonstrate a considerable quality

enhancement compared with the initial ones. Notably, although the Elo rating is not the direct optimization target, its progressive increase emerges from the system’s information feedback loops that enable recursive self-improvement.

To further contextualize this observation, we focused our analysis on a subset of 15 research goals, curated as challenging problems by seven biomedical experts in their respective fields of expertise (Fig. 2b). These experts held doctoral degrees in biological or life science disciplines and were actively working as postdoctoral researchers or faculty members at academic research institutions. These research goals were consistently structured and formatted, encompassing a research title, a clear set of goals, preferences specifying biological or disease areas of interest, desirable solution attributes and constraints on experimental techniques.

In addition to the research goals, the experts provided their ‘best guess’ hypotheses or solutions. We then included outputs from other state-of-the-art LLMs and reasoning models at the time of this study (Gemini 2.0 Pro Experimental, Gemini 2.0 Flash Thinking Experimental 12-19, OpenAI o1, OpenAI o3-mini-high, and DeepSeek R1) in a tournament along with the expert’s best guess and Co-Scientist for each curated goal. Performance was assessed using the Co-Scientist Elo rating metric.

The trends previously observed with test-time compute scaling in Fig. 2a were consistent within this subset. Furthermore, as shown in Fig. 2b, Co-Scientist eventually substantially surpassed the other frontier LLMs and reasoning models in Elo rating with iterative improvement. Notably, newer reasoning models, such as OpenAI o3-mini-high and DeepSeek R1, demonstrated competitive performance while requiring much less compute and reasoning time. Finally, we observed no evidence of performance saturation as measured by Elo, suggesting that further scaling of test-time compute in this paradigm could yield continued improvements in the result quality of Co-Scientist, provided that the research goal is tractable and benefits from the search-and-explore paradigm. It is worth noting again that the Co-Scientist architecture is model agnostic and is likely to benefit from further advancements in frontier and reasoning LLMs, such as the most recent Gemini 3 models.

Building on Co-Scientist’s ability to refine and improve research hypotheses and proposals iteratively, we investigated its potential to improve on expert best-guess solutions. Consistent with our previous observations, Co-Scientist demonstrated the ability to enhance expert best-guess solutions over time, as evidenced by the Elo metric in Extended Data Fig. 1. Notably, the improvement trends initially mirrored those of Co-Scientist’s self-generated solutions but subsequently surpassed them as measured by Elo. While this is a preliminary finding requiring further validation, it suggests a new paradigm of human–AI collaboration in scientific discovery with systems such as Co-Scientist augmenting and accelerating the work of expert scientists.

Co-Scientist yields potentially impactful results for experts
To obtain expert feedback and assess preferences, we conducted a small-scale expert evaluation on 11 of the 15 previously curated

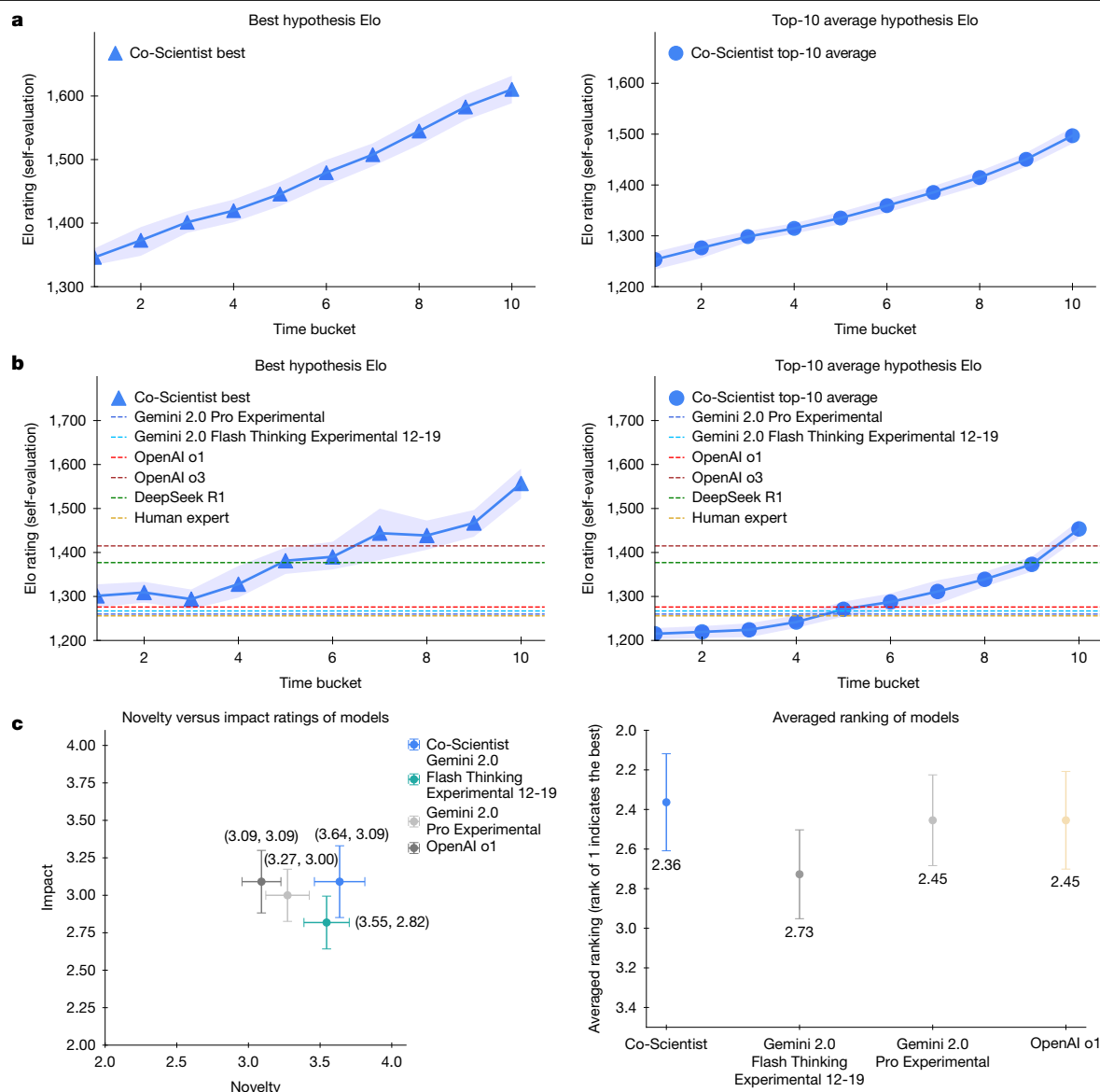


Fig. 2 | Scaling test-time compute enhances Co-Scientist's scientific thinking and hypotheses quality. **a**, Impact of test-time compute scaling on Co-Scientist hypotheses quality measured by Elo auto-evaluation. Generated hypotheses across 203 diverse scientific research goals were partitioned into ten sequential temporal buckets. The continuous upward trend in both the maximum individual Elo rating (left) and the top-ten average Elo rating (right) suggests that the system has the capacity for self-improvement towards high-quality scientific hypotheses for complex problems. **b**, Auto-evaluation benchmarking against standard language models and human experts. Evaluated on a subset of 15 highly complex, expert-curated biomedical goals, Co-Scientist leverages test-time compute scaling to surpass and generate higher-quality and

novel hypotheses compared with human domain experts and state-of-the-art large language and reasoning models (for example, OpenAI o1, o3-mini-high, DeepSeek-R1). **c**, Blinded human expert evaluation: independent domain experts rigorously assessed AI system hypotheses across 11 curated open biomedical problems. Left, the average expert ratings for hypothesis novelty and impact, evaluated independently on a five-point Likert scale and presented as grouped bar charts. Right, the overall expert preference ranking. Co-Scientist consistently achieved the highest expert ratings in novelty and impact and was selected as the preferred AI system by experts. For **a** and **b**, data are mean $\pm$ 95% confidence intervals. For **c**, data are mean $\pm$ probable errors. The exact sample sizes (n) are 203 research goals (**a**), 15 goals (**b**) and 11 goals (**c**).

research goals. We asked the experts who curated the research goals to assess outputs from Co-Scientist, Gemini 2.0 Flash Thinking Experimental 12-19, Gemini 2.0 Pro Experimental, and OpenAI o1 models. Specifically, they provided a preference ranking (1 being most preferred and 4 being least preferred) and rated the novelty and impact of the proposed solutions on a five-point scale, ranging from 1 (worst) to 5 (best) following this rubric. Novelty: higher-ranked outputs should propose hypotheses that, to the best of the expert's knowledge, have not been previously published in any form. Hypotheses similar to existing proposals, even with minor modifications, should rank lower and exact replicas of previously proposed and performed experiments

should receive the lowest ranking. Impact: higher-ranked outputs should address important open questions in the field and have the potential to substantially advance scientific understanding or lead to practical applications.

Across 11 expert-evaluated research goals, the outputs generated by Co-Scientist were most preferred and rated higher in novelty and impact axes compared with the other baseline models. Specifically, Co-Scientist received an average preference rank of 2.36, and novelty and impact ratings of 3.64 and 3.09 (out of 5) as shown in Fig. 2c. These evaluations reflect subjective expert assessments, not objective ground truth. Notably, the human expert preferences also appear

to be concordant with relative Elo ratings as can be inferred from Fig. 2b,c.

We also conducted the preference ranking evaluation for the 15 goals between Co-Scientist and other LLM and reasoning model baselines using OpenAI o3-mini-2025-01-31, o1-preview-2024-09-12, Gemini 2.0 Pro Experimental and Gemini 2.0 Flash Thinking Experimental 01-21 as judges (LLM-as-a-judge evaluation). Co-Scientist outputs were the most preferred by all four evaluation judge LLMs as shown in Extended Data Fig. 2. Owing to the small scale of these evaluations, further studies are necessary for any reliable conclusions. We present a more comprehensive clinical expert evaluation focused on Co-Scientist proposals for drug repurposing in Supplementary Note 4 and Supplementary Fig. 7.

Real-world validations

Drug repurposing with Co-Scientist

Rigorous assessment of a system's ability to generate novel hypotheses for complex scientific problems necessitates end-to-end experimental validation. However, due to the challenging, time-consuming and resource-intensive nature of such endeavours, large-scale experimental validation is infeasible. Instead, we selected areas of unmet clinical need to serve as a strong benchmark for the end-to-end system evaluation of the system's hypotheses-generation ability. Importantly, all experimental validations were conducted in collaboration with expert scientists, who provided guidance to Co-Scientist and prioritized wet-laboratory experiments.

Our first end-to-end validation of Co-Scientist is in drug repurposing, for which the goal was to identify novel therapeutic indications for existing, approved drugs beyond their original use. This approach can accelerate the identification of treatments for complex and rare diseases, as repurposed drugs have established safety profiles and are readily available. From a technical standpoint, this is a complex search-and-explore problem involving a large but finite set of drug-disease pairs as noted in Table 1.

Given Co-Scientist's ability to synthesize and integrate information across a vast body of scientific and clinical literature, we hypothesized that drug repurposing would be an ideal test of the system's abilities. The validation of Co-Scientist's predictions was performed using a multifaceted approach, incorporating computational biology analyses, oncologist expert feedback and *in vitro* wet-laboratory experiments using cancer cell lines.

We constrained Co-Scientist to explore potential repurposing hypotheses from a curated list of 2,300 approved drugs across 34 cancer types and conducted an oncologist expert review of the predictions (Supplementary Note 5.1 and Supplementary Fig. 8). Building on the positive feedback from clinical experts, we conducted *in vitro* wet-laboratory validation experiments for drug-repurposing hypotheses generated by Co-Scientist for acute myeloid leukaemia (AML), an aggressive haematological malignancy marked by uncontrolled proliferation of myeloblasts, resulting in impaired haematopoiesis. There is a critical unmet need for effective therapeutic options in the context of disease recurrence²². The cell-line-based experiments conducted here serve as an initial biological validation step for Co-Scientist hypotheses, with intentionally straightforward methodology following established protocols. We selected four AML cell lines for covering different AML subtypes (MOLM-13, KG-1a, HL-60 and NOMO-1) and a non-AML cell line (TK6), on the basis of the rationale provided by Co-Scientist and the clinical expert in the loop (Supplementary Note 5.2). It is important to emphasize that these wet-laboratory experiments function as a viability check of the drug-repurposing pipeline, yet they are not a replacement for the rigorous preclinical and clinical assessment that is typically required for therapeutic validation. They provide an efficient biological reality check enabling us to rapidly evaluate AI-generated hypotheses before committing to more resource-intensive validation studies necessary for clinical translation.

Wet-laboratory validation of Co-Scientist AML drug-repurposing candidates

The candidate selection for wet-laboratory experiments was performed with meticulous expert oversight. Thirty top-ranked drug candidate hypotheses were shared with expert oncologists (an example detailed Co-Scientist output is provided in Supplementary Note 6). The experts evaluated the hypotheses, selecting drug candidates on the basis of their potential to modulate key molecular signalling pathways associated with disease progression and resistance.

The primary selection criteria prioritized compounds with multipathway activity, specifically those targeting dysregulated inflammatory signalling, metabolic reprogramming and aberrant cell proliferation. Emerging evidence indicates that these interconnected biological processes have critical roles in AML relapse and treatment resistance²³. Candidates were further prioritized on the basis of preclinical mechanistic insights demonstrating relevance to AML biology, including their predicted effects on leukaemic cell survival, interactions within the bone marrow microenvironment and mechanisms underlying therapeutic resistance.

On the basis of potential mechanisms of action, five initial drug-repurposing candidates—binimetinib, pacritinib, cerivastatin, pravastatin and dimethyl fumarate—were selected for further wet-laboratory validation in AML. A list of drug details is provided in Supplementary Note 5.3.

Of the five drugs tested (the experimental setup is described in Supplementary Note 5.4), binimetinib, pacritinib and cerivastatin demonstrated inhibition of cell viability (Fig. 3a–c). Notably, binimetinib, which is already approved for the treatment of metastatic melanoma, exhibited an half-maximal inhibitory concentration (IC₅₀) as low as 2 nM in all AML cell lines (except for NOMO-1), but this was much higher in the TK6 non-AML cell line (Fig. 3a and Extended Data Fig. 3). While binimetinib demonstrated limited efficacy as monotherapy in heavily pretreated, RAS-mutant relapsed/refractory AML in a previous phase II study, Co-Scientist proposed an alternative repurposing strategy for frontline treatment without molecular profiling, on the basis of modulation of baseline RAS–MEK–ERK pathway activity that may influence sensitivity to conventional chemotherapy in treatment-naïve disease²⁴. This result shows the promise of Co-Scientist to aid with drug repurposing.

Single-agent AML repurposing candidates

We next investigated Co-Scientist's ability to autonomously propose single-agent drug-repurposing candidates without oversight. Towards this, the system was directed to generate a ranked list of repurposing candidates for AML that were not previously repurposed for the target indication and without any previous preclinical evidence. Furthermore, Co-Scientist did not receive any additional external inputs, such as the DepMap scores or human expert feedback. We then determined whether these candidates suggested by Co-Scientist could be validated in the laboratory.

For *in vitro* laboratory validation, the domain experts reviewed and selected the top three repurposing candidates to treat AML: nanvuranal, KIRA6 and leflunomide.

The detailed Co-Scientist output, including the hypothesis, rationale and self-generated novelty review, is provided for KIRA6 in Supplementary Note 6. As can be seen, the system identifies that targeting IRE1α in the context of AML has been explored²⁵ previously but not with the specific drug KIRA6, suggesting that the system is reasonably well calibrated in its assessment of novelty. The system suggests an overall moderate level of novelty for the hypothesis.

Of the three drugs tested, treatment with the IRE1α inhibitor KIRA6 showed inhibition of cell viability in several AML cell lines representing different molecular subtypes, KG-1a, MOLM-13, HL-60 and NOMO-1, and the non-AML cell line TK6 as a control (Fig. 3d–h, Supplementary Note 5.2 and Supplementary Table 4). The IC₅₀ values of KIRA6 were all

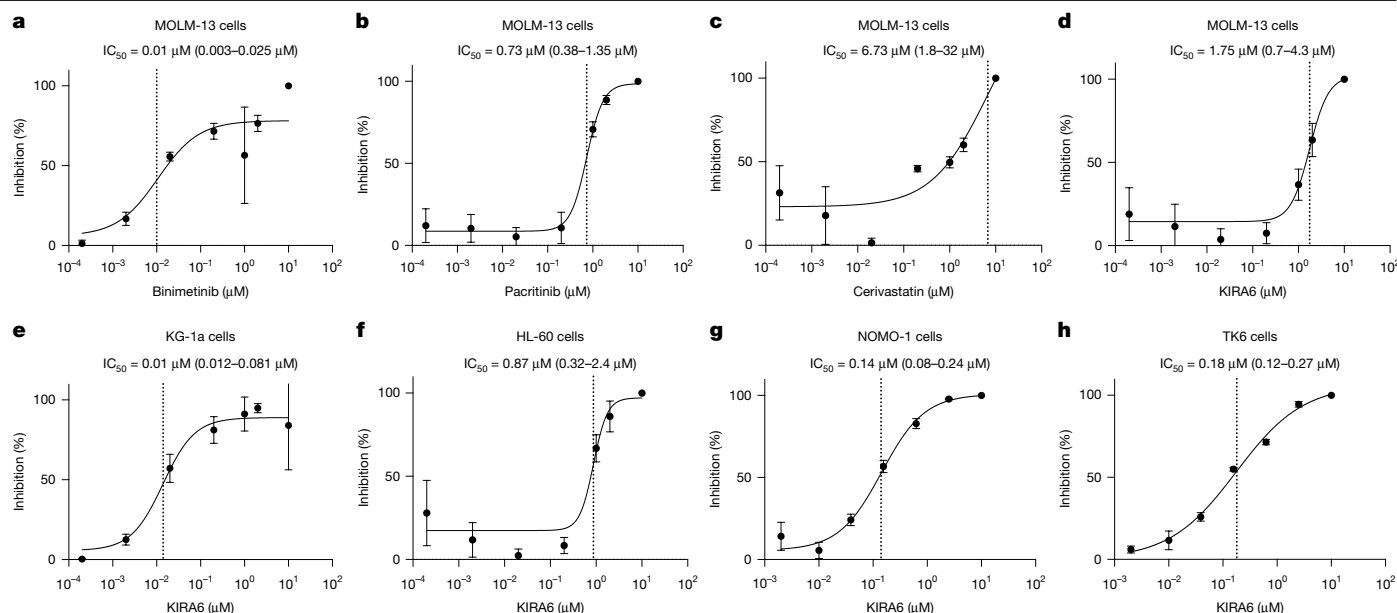


Fig. 3 | In vitro biological validation of Co-Scientist generated single-agent repurposing candidates for AML. Co-Scientist identified promising biologically active compounds, ranging from candidates with existing preclinical rationales to completely novel therapeutic targets for AML. **a–c**, Dose–response curve of MOLM-13 AML cells treated with binimetinib (**a**), pacritinib (**b**) or cerivastatin (**c**). Co-Scientist nominated candidates with existing evidence, demonstrating potent anti-leukaemic activity. **d–h**, Dose–response curves of the completely novel Co-Scientist-predicted candidate KIRA6 (an IRE1α inhibitor) evaluated in different AML cell lines (MOLM-13 (**d**), KG-1a (**e**), HL-60 (**f**) and NOMO-1 (**g**)) and in the normal lymphoblastoid control

cell line TK6 (**h**). KIRA6 exhibits highly selective cytotoxicity against the KG-1a AML cell line compared with the non-malignant TK6 control. The 18-fold separation establishes a promising in vitro therapeutic window and suggests Co-Scientist's promising capability to search, reason and identify biologically active compounds. The x axis represents drug concentration (μM) on a logarithmic scale, and the y axis represents the percentage of growth inhibition. Data are mean ± s.d. of $n = 3$ biologically independent experiments. Exact IC_{50} values were determined using nonlinear regression curve fitting; numbers in parentheses above plots indicate 95% confidence interval.

in the nanomolar or low micromolar range, but it was much more effective in KG-1a cells (IC_{50} of 10 nM) than in the non-AML control cell line TK6 (IC_{50} of 180 nM). KIRA6 was also slightly more effective in NOMO-1 cells, with an IC_{50} of 144 nM but, notably, MOLM-13 and HL-60 cells were markedly less sensitive to KIRA6 (IC_{50} values of 1,750 nM and 870 nM, respectively). The 18-fold separation between the highly primitive KG-1a cells and the normal lymphoblastoid TK6 line highlights a potential selective therapeutic window. The differential sensitivities observed across distinct AML subtypes indicate that IRE1α blockade may be most effective in targeting primitive, stem-like AML populations over more differentiated lineages. Comprehensive cytogenetic rationales and molecular mechanisms correlating these varying cell line sensitivities to the IRE1α–XBPI axis are described in Supplementary Note 5.2. Nanvuranlat and leflunomide instead show limited effect on MOLM-13 cells (Extended Data Fig. 4).

Synergistic AML drug combinations

A common strategy for more effective treatment is the combination of drugs that synergistically target different disease pathways, but searching and screening for these combinations becomes exponentially more complex as the number of drugs increases. This is a complex task even for experts; at the same time, it may be well suited for AI. To investigate this, we used Co-Scientist to identify promising synergistic multi-drug regimens for AML. We then evaluated seven drug combinations proposed by Co-Scientist in the MOLM-13 and KG-1a cell lines. In MOLM-13 cells, responses were predominantly synergistic across both dual (for example, JNJ-64619178 + selinexor) and triple combinations (such as JQ1 + olaparib + MSA2). By contrast, KG-1a cells exhibited highly context-dependent responses with a mixture of synergy and antagonism, probably reflecting their distinct chemoresistant molecular profile (*TP53* mutant). A comprehensive summary of all interaction patterns is provided in Fig. 4, Extended Data Figs. 5 and 6, Extended Data

Table 2 and Supplementary Note 5.5. These patterns probably reflect the underlying molecular profiles of the two cell lines (Supplementary Notes 5.3 and 5.5). Further mechanistic studies will be required to define the molecular determinants of response to combination therapy across AML subtypes, and to identify predictive biomarkers that could enable rational regimen selection.

These results not only demonstrate Co-Scientist's ability to identify potent single agents but also highlight its use in proposing new synergistic drug combinations that have potential in addressing therapeutic resistance and treatment-refractory disease. Importantly, it shows the potential of Co-Scientist to identify new drug combinations without the requirement of large-scale wet-laboratory screening, which becomes exponentially more costly and difficult as larger combinations are considered. Thus, Co-Scientist may enable research that was previously restricted by wet-laboratory design and feasibility.

Guiding clinical translation design

While the results are promising, translating these predictions from Co-Scientist into clinical practice will be highly challenging, as the complexity of a disease model, patient heterogeneity and disease variability cannot be fully captured in such limited in vitro experiments. Even if a hypothesis generated by Co-Scientist is well reviewed by oncologists and supported by preclinical rationale and strong in vitro experiments, this does not guarantee in vivo efficacy or clinical success. Factors such as drug bioavailability, pharmacokinetics, off-target effects and patient selection criteria can all impact onward clinical trial outcomes. Moreover, in case of haematological malignancies, the tumour microenvironment and systemic interactions may introduce unforeseen resistance mechanisms, further complicating translation from hypothesis to clinical benefit.

To more faithfully approximate the parameters that govern real-world therapeutic decision-making, we tasked Co-Scientist with a

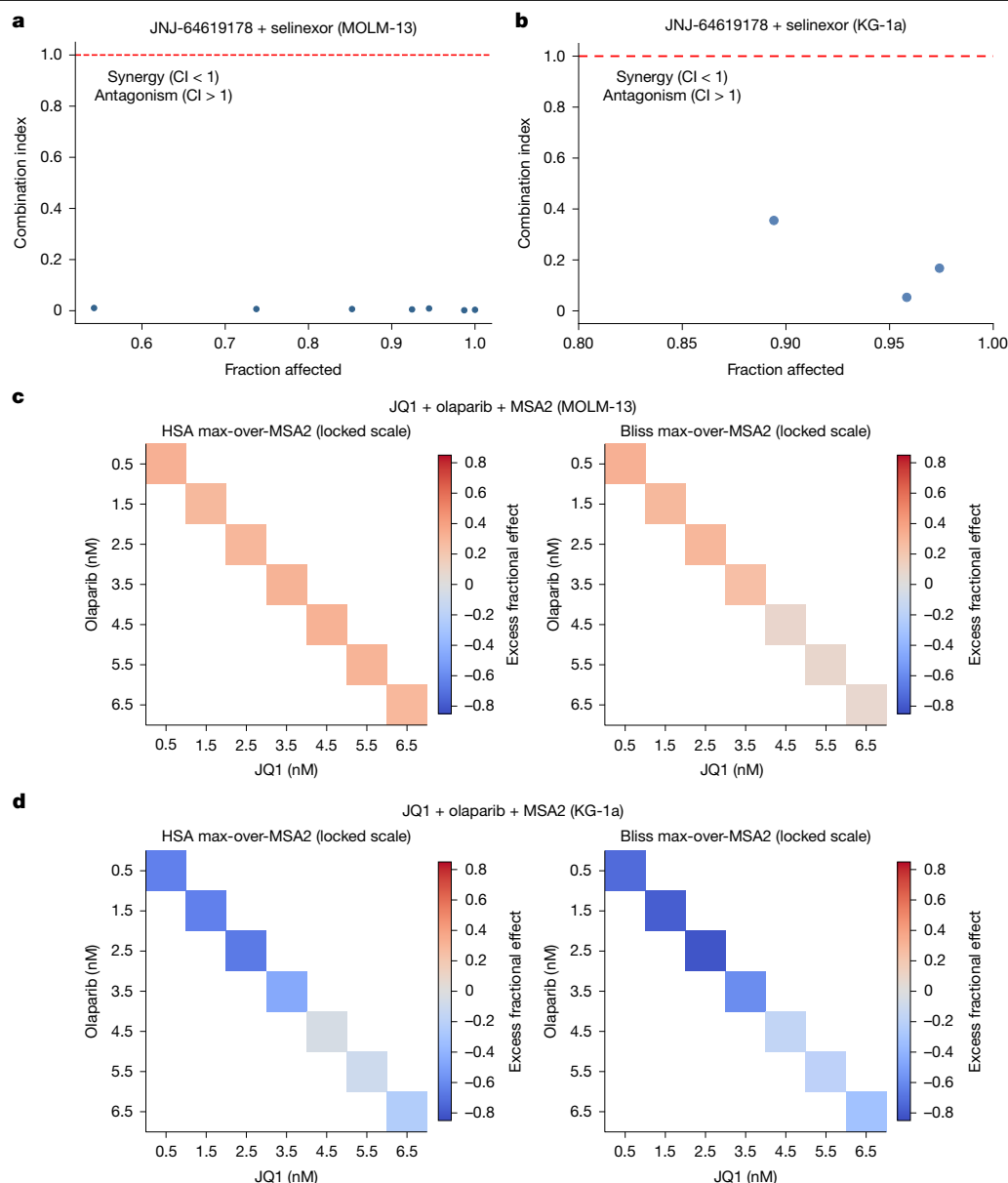


Fig. 4 | Validation of Co-Scientist predicted synergistic multi-drug combinations for AML. Co-Scientist successfully navigated high-dimensional combinatorial spaces to propose effective multi-drug therapy regimens, validated here in the AML cell lines MOLM-13 (**a** and **c**) and KG-1a (**b** and **d**). **a,b**, Quantitative synergy analysis of the dual combination JNJ-64619178 and selinexor. The plot illustrates the relationship between the combination index (CI) and the fraction affected (Fa) using the Chou–Talalay method. The horizontal red dashed line represents strictly additive effects (CI = 1.0). Datapoints falling below this threshold denote the synergy zone (CI < 1), while points above denote the antagonism zone (CI > 1), confirming strong synergistic interactions for this Co-Scientist-proposed double-drug

combination. **c,d**, Excess fractional effect heat map for the triple-drug combination JQ1, olaparib and MSA2. Synergy is quantified across a matrix of drug concentrations (nM) using the highest single agent (HSA) and Bliss independence models. The colour scale illustrates the deviation from predicted additive effects: the red regions indicate a positive excess effect (synergy), and the blue regions represent a negative excess effect (antagonism). These results suggest Co-Scientist's ability to identify highly active, complex combinatorial treatments without exhaustive empirical screening. Complete interaction profiles are available in Extended Data Figs. 5 and 6 and Extended Data Tables 1 and 2. For all synergy analyses, experiments were performed in $n = 3$ biologically independent replicates.

structured translational analysis using a detailed clinical variable framework encompassing patient demographics, ELN2022 risk stratification, molecular features, preclinical activity and established safety and pharmacokinetic/pharmacodynamic data (Supplementary Note 5.6).

Here, for example, Co-Scientist's structured translational analysis successfully identified a specific clinical niche for binimetinib: frail, heavily pretreated patients with AML. The system accurately deduced that binimetinib's unique metabolic pathway (UGT1A1) circumvents severe CYP3A4-dependent drug–drug interactions common with azole antifungals—a major limitation for current targeted therapies

(Supplementary Note 5.6). Extended results and methodological details are provided in Supplementary Note 5.2 and 5.6. Taken together, these analyses illustrate how Co-Scientist can help clinician scientists move beyond initial hypothesis generation to synthesize diverse clinical and biological variables into testable, clinically grounded therapeutic strategies.

Identifying therapeutic targets for liver fibrosis

Co-Scientist used a method employing human hepatic organoids coupled with live-cell imaging to identify novel therapeutic targets

for severe liver fibrosis^{1,26,27}. Co-Scientist was tasked with generating hypotheses on target epigenetic alterations (three top-ranked hypotheses were selected by experts) and identifying drugs targeting these predicted epigenetic modifiers. It successfully identified three epigenetic modifiers and drugs targeting them, and two of them exhibited significant anti-fibrotic activity in the hepatic organoids without causing cellular toxicity. Critically, one of the effective drugs (vorinostat) is already approved by the US Food and Drug Administration (FDA) for another cancer indication, creating an opportunity for drug repurposing for liver fibrosis treatment¹. This example also highlights the potential of AI systems such as Co-Scientist to make unexpected connections across disparate disciplines and diseases (cancer and liver fibrosis) and synthesize helpful and impactful hypotheses and discoveries.

Recapitulating a breakthrough in AMR

Co-Scientist was also challenged to independently identify the mechanism behind the broad host range and rapid spread of cf-PICs, mobile elements that carry virulence and antibiotic-resistance genes across diverse bacterial species (including *Escherichia coli* and *Klebsiella pneumoniae*). With only minimal background information, Co-Scientist independently and accurately proposed the top-ranked hypothesis that cf-PICs interact with diverse phage tails to expand their host range². This finding, generated by the AI in just 2 days, precisely matched the primary discovery of an independent, co-timed genomic and experimental study before completing peer review²¹. This convergence and recapitulation is another demonstration of Co-Scientist's potential to accelerate scientific discovery by synthesizing complex scientific information and generating rigorous scientific hypotheses on-par with experts.

Discussion

Here we report the development and initial validation of a multi-agent Gemini-based AI system, Co-Scientist, designed as a structured scientific thinking engine to accelerate novel scientific discovery. Co-Scientist moves beyond conventional computational approaches through the *in silico* implementation of a multi-agent architecture that mirrors the core aspects of the scientific method. Instead of brute-force generation, the system iteratively refines hypotheses through a 'generate, debate, evolve' paradigm. This method, which incorporates self-debate, tournament-based selection and iterative evolution and refinement, enables a progressive convergence on high-quality, well-supported hypotheses, thereby scaling research ideation with test-time compute rather than exhaustive generation. The system's context memory, combined with the iterative self-improvement cycle, functions as an emergent internal model of the scientific research process. While not an explicit symbolic model, it represents a progressively more coherent and interconnected state of knowledge, facilitating the synthesis of information and the identification of knowledge gaps.

The practical use of this approach was demonstrated through the generation of novel and experimentally tractable hypotheses across three challenging and varied biomedical problems. In oncology, Co-Scientist identified drug-repurposing candidates for AML that showed *in vitro* efficacy at clinically relevant concentrations. For liver fibrosis, it proposed new epigenetic targets, leading to the experimental validation of several anti-fibrotic compounds, including one FDA-approved drug. Furthermore, in microbiology, the system independently recapitulated an at-the-time unpublished mechanism of mobile genetic element transfer between bacteria. These findings provide preliminary evidence that Co-Scientist can contribute meaningfully to scientific discovery by amplifying scientists.

This system's architecture is model agnostic, enabling it to take advantage of the advancing capabilities of frontier LLMs without requiring retraining of the whole agentic framework, such as Gemini 3, GPT 5.4 and Opus 4.6. With the latest advances in frontier models,

we expect further substantial improvements in the quality of hypotheses generated and the complexity of scientific tasks that the system can autonomously accomplish.

Despite these promising early results, several limitations must be addressed. Co-Scientist's knowledge is constrained by its reliance on open-access scientific literature, which may lead to the omission of critical previous work behind paywalls and a systemic lack of access to negative experimental results. Furthermore, the quality of generated hypotheses relies on the mixed and contradictory quality of the source literature; there is therefore a risk of propagating erroneous or irreproducible findings. A key future direction is the development of agents with enhanced provenance capabilities to trace claims to specific figures or data within a source, mitigating the impact of unreliable literature.

Co-Scientist also inherits the intrinsic limitations of its underlying models, including imperfect factuality and the potential for hallucinations. Improving reasoning capabilities is a critical area for future work. Moreover, the validation of Co-Scientist's hypotheses, while successful, remains preliminary.

Finally, the broader integration of such AI systems into the scientific workflow requires careful consideration of potential bias, which could risk diminishing critical thinking or homogenizing research directions. While AI has the potential to democratize access to scientific information, particularly in resource-limited settings, it is essential to develop robust verification methods and maintain rigorous peer review to ensure that AI can augment, rather than replace, human scientific reasoning and creativity. Improper use of such AI systems without rigorous peer review and guardrails could also lead to worsening of the scientific reproducibility crisis through production of low-quality scientific artifacts. Further details regarding safety and ethical implications are provided in Supplementary Note 7.

The continued development of Co-Scientist will focus on three key areas. Immediate improvements will target the system's robustness by enhancing learning and knowledge base, literature search capabilities to broaden access, implementing more rigorous fact-checking against external databases and tools, and improving citation recall. Future advancements will focus on expanding the system's core capabilities. This includes integrating agents that can directly reason over public databases and multimodal data, enabling bioinformatics and data science tasks. The implementation of reinforcement learning from human and experimental feedback could further optimize the hypothesis generation and refinement process.

Expanded evaluations are also necessary to assess Co-Scientist's generalizability across a wider range of scientific disciplines. This requires developing more objective and automated evaluation metrics that move beyond current ranking systems and engaging a larger cohort of domain experts to stress test the system with diverse and complex research queries.

In the fullness of time, integrating Co-Scientist with laboratory automation platforms could create a closed-loop, autonomous system for hypothesis generation, experimental validation and iterative learning, substantially accelerating the pace of scientific discovery.

Conclusion

Co-Scientist represents a promising step towards AI-assisted augmentation of scientists and acceleration of scientific discovery. Its ability to think scientifically, generate testable hypotheses across diverse scientific and biomedical domains, some supported by experimental findings, along with the capacity for recursive self-improvement with increasing compute, demonstrates the promise of meaningfully accelerating scientists' endeavours to resolve grand challenges in human health, medicine and science. This innovation opens numerous questions and opportunities. Applying the empiric and responsible approach of science to Co-Scientist itself can thereby enable safe

exploration of its undoubted potential, including how collaborative and human-centred AI systems might be able to augment human ingenuity and accelerate scientific discovery.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-026-10644-y>.

- Guan, Y. et al. AI-assisted drug re-purposing for human liver fibrosis. *Adv. Sci.* **12**, e08751 (2025).
- Penadés, J. R. et al. AI mirrors experimental science to uncover a novel mechanism of gene transfer crucial to bacterial evolution. *Cell* **188**, 6654–6665 (2025).
- Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
- Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl Acad. Sci. USA* **79**, 2554–2558 (1982).
- Hinton, G. E. & Sejnowski, T. J. in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* Vol. 1 (eds Feldman, J. A. et al.) 282–317 (MIT Press, 1986).
- Guo, D. et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633–638 (2025).
- Jaech, A. et al. OpenAI O1 system card. Preprint at <https://arxiv.org/abs/2412.16720> (2024).
- Gemini Team, Google. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. Preprint at <https://arxiv.org/abs/2403.05530> (2024).
- Wiesinger, J., Marlow, P. & Vuskovic, V. Agents. *kaggle* <https://www.kaggle.com/whitepaper-agents> (2024).
- Hinton, G. E., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. Preprint at <https://arxiv.org/abs/1503.02531> (2015).
- Gemma Team, Google. Gemma: open models based on gemini research and technology. Preprint at <https://arxiv.org/abs/2403.08295> (2024).
- Leslie, D. et al. ‘Frontier AI’, power, and the public interest: who benefits, who decides? *Harvard Data Sci. Rev.* <https://doi.org/10.1162/99608f92.4a42495c> (2024).
- Chen, L. et al. Are more LLM calls all you need? Towards the scaling properties of compound AI systems. *Adv. Neural Inform. Process. Syst.* **37**, 45767–45790 (2024).
- Comanici, G. et al. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Preprint at <https://arxiv.org/abs/2507.06261> (2025).
- Gower, B. *Scientific Method: a Historical and Philosophical Introduction* (Routledge, 2012).
- Snell, C., Lee, J., Xu, K. & Kumar, A. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)* (2025).
- Brown, N. & Sandholm, T. Superhuman AI for multiplayer poker. *Science* **365**, 885–890 (2019).
- Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
- Ringel, M. S., Scannell, J. W., Baedeker, M. & Schulze, U. Breaking Eroom’s law. *Nat. Rev. Drug Discov.* **19**, 833–834 (2020).
- Pushpakom, S. et al. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **18**, 41–58 (2019).
- He, L. et al. Chimeric infective particles expand species boundaries in phage-inducible chromosomal island mobilization. *Cell* **188**, 6636–6653 (2025).
- Döhner, H., Weisdorf, D. J. & Bloomfield, C. D. Acute myeloid leukemia. *N. Engl. J. Med.* **373**, 1136–1152 (2015).
- Guo, Q. et al. NF-κB in biology and targeted therapy: new insights and translational implications. *Signal Transduc. Target. Ther.* **9**, 53 (2024).
- Maiti, A. et al. Phase II trial of MEK inhibitor binimetinib (MEK162) in RAS-mutant acute myeloid leukemia. *Clin. Lymphoma Myeloma Leuk.* **19**, 142–148 (2019).
- Philippe, C. et al. Pivotal role of the endoplasmic reticulum stress-related XBP1s/miR-22/SIRT1 axis in acute myeloid leukemia apoptosis and response to chemotherapy. *Leukemia* **38**, 1764–1776 (2024).
- Guan, Y. et al. A human multi-lineage hepatic organoid model for liver fibrosis. *Nat. Commun.* **12**, 6138 (2021).
- Guan, Y. et al. Live-cell imaging of human liver fibrosis using hepatic micro-organoids. *JCI Insight* **10**, e187099 (2024).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026

Methods

Overview of Co-Scientist architecture

Co-Scientist uses a multi-agent architecture built on Google's Gemini. In this study, we used Gemini 2.0 models as the base foundational LLM for all agents¹⁴, integrated within an asynchronous task execution framework. This framework allows flexible scaling of test-time compute resources, facilitating advanced scientific thinking and reasoning. Given a research goal specified by an expert scientist in natural language, Co-Scientist generates hypotheses that adhere to the following default criteria. (1) Alignment with the provided research goal. The generated outputs must precisely align with the research goals, preferences and constraints defined by the scientist. (2) Plausibility. The system outputs should be free of readily apparent flaws. Any potential contradictions with previous literature or established knowledge must be explicitly stated and justified. (3) Novelty. A key objective of Co-Scientist is to generate novel hypotheses, conjectures and research plans grounded in previous literature, rather than simply synthesizing existing information (a capability already addressed by existing 'deep research' tools²⁸). (4) Testability. The system outputs should be amenable to empirical validation within the constraints specified by the scientist. (5) Safety. The system outputs will be controlled to prevent enabling unsafe, unethical or harmful research. Aside from these default criteria, Co-Scientist can be configured with additional criteria, preferences and constraints as needed. For example, it can be configured to generate outputs in formats preferred by the researcher to improve interpretability and readability.

At a high level, Co-Scientist comprises four key components. (1) A natural-language input-output interface. Scientists interact with and supervise the system primarily through natural language. This allows them to not only define the initial research goal but also refine it at any time, provide feedback on generated hypotheses (including their own solutions) and generally steer and guide the system's progress. (2) Asynchronous task framework. Co-Scientist uses a multi-agent system in which specialized agents operate as worker processes within an asynchronous, continuous and configurable task execution framework. A dedicated Supervisor agent manages the worker task queue, assigns specialized agents to these processes and allocates resources. This design enables the system to flexibly and effectively use computational resources and iteratively improve its scientific reasoning and quality of hypotheses. (3) Specialized agents. After accounting for inductive biases and scientific priors derived from the scientific method, the process of scientific reasoning and hypothesis generation is broken down into subtasks. Individual, specialized agents, each equipped with customized instruction prompts, are designed to execute these subtasks. These agents operate as workers coordinated by the Supervisor agent. (4) Context memory. To enable iterative computation and scientific reasoning over long time horizons, Co-Scientist uses a persistent context memory to store and retrieve states of the agents and the system during the course of the computation. The specific Co-Scientist design was arrived at with iterative developments and feedback from expert scientists and is reflective of the current capabilities of the underlying LLMs. The Co-Scientist multi-agent architecture is depicted and summarized in Fig. 1b.

Throughout the next section, we use a recurring example: generating hypotheses for exploring the biological mechanisms of amyotrophic lateral sclerosis (ALS) to illustrate the various components of Co-Scientist. Although this example has been reviewed by domain experts, it remains illustrative and may contain errors. Importantly, this example does not aim to suggest potential therapeutic avenues for ALS and should be interpreted with utmost caution. We have also provided the pseudocode demonstrating agent logic in Supplementary Note 8. All of the prompts used in the agents are listed in Supplementary Note 9, and all the examples are listed in the Supplementary Note 10.

From research goal to research plan configuration

The research goal, specified by the scientist, serves as the entry point to Co-Scientist. Leveraging the multimodal and long context capabilities of Gemini models, Co-Scientist efficiently processes research goals of varying complexity, from simple statements to extensive documents spanning tens of thousands of natural language tokens or other relevant data (for example, including hundreds of previous publication PDF files). The research goal may also incorporate specific constraints, attributes and preferences related to the scientist's particular laboratory setting or field of work.

Co-Scientist then parses the goal to derive a research plan configuration for generating research proposals. This configuration captures the desired proposal preferences, attributes and constraints. For example, it specifies whether Co-Scientist should exclusively propose novel hypotheses. It also specifies the criteria for evaluating hypothesis quality, such as novelty and experimental feasibility. These criteria are then used by the system during its auto-evaluation, tournament debates and self-improvement phases. The attributes, preferences and evaluation criteria can all be customized to a given research goal. To illustrate this process, we present an example research goal and its corresponding parsed research plan configuration in Supplementary Note 10.1, in which the goal is to develop a novel hypothesis related to phosphorylation of the nuclear pore complex as a causative mechanism for ALS²⁹.

On the basis of the research plan configuration, the Supervisor agent initiates the creation of a task queue and begins orchestrating the specialized agents. The system operates continuously and asynchronously. Periodically, the Supervisor agent calculates a comprehensive set of summary statistics, reflecting the system's state and progress toward the specified research goal. These statistics inform decisions regarding resource allocation and the determination of whether a terminal state for the overall computation has been reached. The state is periodically written to the associated context memory of the system and leveraged as feedback in subsequent rounds of computation. It also enables easy restarts in case of any failure in the system components.

The specialized agents underpinning Co-Scientist

At the heart of Co-Scientist is a coalition of specialized agents, each orchestrated by the Supervisor agent. These agents are designed to emulate the scientific reasoning process, enabling them to generate novel hypotheses and research plans. Each agent is provided a library of strategies (that is, a collection of prompts) to better explain and help it perform its assigned task. They are also equipped to interact with and utilize external tools, such as web search engines and specialized AI models, through application programming interfaces (APIs). These specialized agents are enumerated below: (1) Generation agent. The agent initiates the research process by generating the initial focus areas, iteratively extending them and generating a set of initial hypotheses and proposals that address the research goal. This involves exploring relevant literature using web search, synthesizing existing findings into new directions and engaging in simulated scientific debates for iterative improvement. (2) Reflection agent. This agent simulates the role of a scientific peer reviewer, critically examining the correctness, quality and novelty of the generated hypotheses and research proposals. Furthermore, it evaluates the potential of each hypothesis to provide an improved explanation for existing research observations (identified through literature search and review), particularly those that may be under-explained. (3) Ranking agent. An important abstraction in Co-Scientist is the notion of a tournament in which different research proposals are evaluated, debated and ranked enabling iterative improvements. The Ranking agent orchestrates an Elo-based tournament³⁰ to assess and prioritize the generated hypotheses at any given time. This involves pairwise comparisons, facilitated by simulated scientific debates, which allow nuanced evaluation of the relative merits of each hypothesis. (4) Proximity agent. This agent asynchronously

Article

computes a proximity graph for generated hypotheses, enabling clustering of similar ideas, de-duplication and efficient exploration of the hypothesis landscape. (5) Evolution agent. Co-Scientist's iterative improvement capability relies heavily on this agent, which continuously refines the top-ranked hypotheses emerging from the tournament. Its refinement strategies include synthesizing existing ideas, using analogies, leveraging literature for supporting details, exploring unconventional reasoning and simplifying concepts for clarity. (6) Meta-review agent. This agent also enables Co-Scientist's continuous improvement by synthesizing insights from all reviews, identifying recurring patterns in tournament debates, and using these findings to optimize other agents' performance in subsequent iterations. This also enhances the quality and relevance of generated hypotheses and reviews in subsequent iterations. At the end of the Co-Scientist computation process, this agent also synthesizes top-ranked hypotheses and reviews into a comprehensive research overview for review by the scientist.

The Supervisor agent's seamless orchestration of these specialized agents enables the development of valid, novel and testable hypotheses and research plans tailored to the input research goal provided by the scientist.

In summary, the Generation agent curates an initial list of research hypotheses satisfying a research goal. These are then reviewed by the Reflection agent and evaluated in a tournament by the Ranking agent orchestrating debate matches involving the hypothesis. The Evolution, Proximity, and Meta-review agents operate on the tournament state to help to improve the quality of the system outputs.

The Supervisor agent periodically computes and writes to the context memory a comprehensive suite of statistics, including the number of hypotheses generated and requiring review, and the progress of the tournament. These statistics also include analyses of the effectiveness of different hypothesis generation methodologies (for example, generating new ideas through the Generation agent versus improving existing ideas through the Evolution agent). On the basis of these statistics, the Supervisor agent then orchestrates subsequent system operations, that is, generating new hypotheses, reviews, tournaments and improvements to existing hypotheses, by strategically weighting and sampling the specialized agents for execution through the worker processes.

Importantly, the Meta-review agent enables feedback propagation and learning without back-propagation techniques (for example, fine-tuning or reinforcement learning)³¹. The Meta-review agent generates feedback applicable to all agents, which is simply appended to their prompts in the next iteration—a capability facilitated by the long-context search and reasoning capabilities of the underlying Gemini models. Through this feedback loop, Co-Scientist continuously learns and improves in subsequent iterations with more compute scaling.

Finally, while our work leverages Gemini 2.0, Co-Scientist framework is model agnostic and portable to other similar models or combinations thereof. Future LLM improvements, such as our Gemini 3 models, will further enhance Co-Scientist's ability to generate scientific hypotheses and perform complex tasks over the longer term.

Generation agent

The Generation agent uses a diverse array of strategies (encoded as prompts), techniques and tools to generate hypotheses, such as the following:

- Literature exploration through web search. The agent iteratively searches the web, retrieves and reads relevant research articles, learns about topics and grounds its reasoning by summarizing previous work and building a knowledge base of scientific facts. It then builds on this summary to generate hypotheses and research plans. An example prompt is given in Supplementary Note 9.1.
- Simulated scientific debates. Here, the Generation agent simulates scientific debates among experts by using self-critique and self-play

techniques. These debates typically involve multiple turns of conversations leading to a refined hypothesis generated at the end. An example prompt is given in Supplementary Note 9.1.

- Iterative assumptions identification. The agent iteratively identifies testable intermediate assumptions, which, if proven true, can lead to scientific discovery. These plausible assumptions and their subassumptions are identified through conditional reasoning hops and subsequently aggregated into complete hypotheses.
- Research expansion. To identify previously unexplored areas of the hypothesis space, the Generation agent reviews existing hypotheses and the research overview and feedback provided by the Meta-review agent in the previous iteration. This is used to inform additional exploration directions in the research hypothesis space.

An example hypothesis and research proposal output from the Generation agent is presented in Supplementary Note 10.2 for the aforementioned research goal regarding explaining a basic mechanism related to ALS. The Generation agent also summarizes and categorizes each generated hypothesis, enabling scientists to quickly grasp the core ideas.

Reflection agent

Reviews are integral to Co-Scientist's effectiveness in generating new proposals. The Reflection agent searches relevant previous work and data (through web search or a dedicated scientist-provided repository), assesses existing experimental evidence for or against a given hypothesis and rigorously verifies the novelty, correctness and quality of generated outputs with tools. Effective reviews filter inaccurate and, when stipulated, non-novel hypotheses. Moreover, they also provide feedback to all of the other agents, driving continuous improvement. The Reflection agent uses the following types of review strategies:

- Initial review. Building on Co-Scientist's default evaluation criteria, the Reflection agent performs an initial review assessing the correctness, quality, novelty and a preliminary assessment of safety (ethics) of the generated hypotheses. A more in-depth discussion on safety considerations is provided in Supplementary Note 7. This initial review, which does not use external tools such as web search, aims to quickly discard flawed, non-novel or otherwise unsuitable hypotheses.
- Full review. If a hypothesis passes the initial review, the Reflection agent performs a full review, using external tools and web searches to identify relevant articles for improved reasoning and grounding. This review evaluates the hypothesis's correctness, quality and novelty similar to the initial review but with full literature search. For correctness and quality, the agent scrutinizes the underlying assumptions and reasoning. For novelty, it summarizes known aspects of the hypothesis and then judges their novelty on the basis of existing literature. An example full novelty review, review critiques and a complete full review example are shown in Supplementary Note 10.3, 10.4 and 10.5, respectively.
- Deep verification review. The Reflection agent also conducts a deep verification review, decomposing the hypothesis into constituent assumptions. Each assumption is further broken down into fundamental subassumptions, decontextualized and independently evaluated for correctness to identify invalidating elements for subsequent filtering. Concurrently, the reasons for potential hypothesis invalidation due to incorrect assumptions are summarized. This deep verification helps Co-Scientist to detect subtle errors within complex hypotheses, such as flaws in reasoning or inaccurate experimental protocols. An identified error does not necessarily invalidate the core hypothesis; the Reflection agent assesses whether the incorrect assumption is fundamental to the hypothesis and incorporates this reasoning into the review. Non-fundamental errors can be addressed during subsequent refinement stages. An example deep verification review is provided in Supplementary Note 10.6 for the previously introduced ALS hypothesis. We also show another example of a deep verification

review through probing questions in the context of drug repurposing for AML in Supplementary Note 10.7.

- **Observation review.** Moreover, the Reflection agent also explores whether a given hypothesis can account for long-tail observations from previous experimental results. This review aims to determine whether the hypothesis can provide insights on existing experimental findings and observed phenomena within relevant articles. For each observation, the agent assesses whether the hypothesis is a superior explanation over existing ones, assuming its validity. Positive observations are summarized and appended to the hypothesis. Note that this review often completes without any important findings (as in the case for the ALS hypothesis example). An example prompt to generate observations is provided in Supplementary Note 9.2. An illustrative example of an observation review is provided in Supplementary Note 10.8 in the context of an alternative hypothesis for explaining a mechanism of AMR.
- **Simulation review.** The Reflection agent also reviews hypotheses by simulating them in a step-wise manner (for example, simulating the mechanism of action or the proposed experiment in the proposal). This simulation enables the agent to identify and summarize potential failure scenarios. This review method is built on the assumption that frontier LLMs may have developed an internal world model of science that enables them to simulate and accurately predict various scientific phenomena.
- **Recurrent/tournament review.** The Reflection agent adapts its full reviews on the basis of Co-Scientist's growing knowledge. By analysing reviewed hypotheses and results of the tournament conducted by the Ranking agent, the Reflection agent identifies recurring issues and improvement opportunities, refining its reviews accordingly.

Moreover, Co-Scientist can incorporate reviews from expert scientists to guide ranking and improvements (further discussed in the 'Expert-in-the-loop interactions with Co-Scientist' section). A key aim is to have the Reflection agent generate a comprehensive set of reviews from multiple angles covering the common methods that scientists use when critiquing and refining research hypotheses and proposals.

Ranking agent

Co-Scientist explores numerous hypotheses and research proposals towards a research goal, necessitating a ranking mechanism to prioritize scientists' time and computational resources toward the most promising candidates. This task is performed by the Ranking agent. The agent orchestrates an Elo-based tournament³⁰ to automatically evaluate and rank all hypotheses, providing supporting rationale. This ranking serves to communicate to scientists an ordered list of research hypotheses and proposals aligned with the research goal. Despite its assumptions and limitations³², Elo remains a good proxy for relative ranking, and it has previously been applied to rank extracted patterns and ideas in games³³. In the future, extensions may be considered³⁴. We set the initial Elo rating of 1,200 for the newly added hypothesis.

As the tournament is computationally intensive, the Ranking agent uses several optimization strategies. Top-ranked hypotheses are compared pairwise in tournament matches through multi-turn scientific debates³⁵. This mitigates ordering bias and focuses on novelty, correctness and testability. Lower-ranked hypotheses undergo single-turn comparisons in a pairwise manner in their tournament match. The agent concludes each comparison with a decision regarding which hypothesis is better. Supplementary Note 9.3 shows example prompts. Supplementary Note 10.9 shows an example of the Ranking agent conducting a scientific debate match in a tournament to compare two hypotheses.

The Ranking agent prioritizes tournament matches as follows: (1) hypotheses are more likely to be compared with similar ones (on the basis of the Proximity agent's graph, described in the next section); (2) newer and top-ranking hypotheses are prioritized for participation in tournament matches. Successful hypotheses quickly achieve

favourable rankings and this informs the tournament state for subsequent iterations.

Proximity agent

The Proximity agent calculates the similarity between research hypotheses and proposals, and builds a proximity graph, taking into account the specific research goal. Although it does not directly participate in hypothesis generation, the Proximity agent assists the Ranking agent in organizing tournament matches and showcasing a diverse range of ideas related to the research goal. This enables scientists to quickly explore areas of interest and easily identify related concepts.

Evolution agent

The Evolution agent continuously refines and improves existing hypotheses and proposals using several strategies, including the following:

- **Enhancement through grounding.** Here the agent attempts to improve hypotheses by identifying weaknesses, generating search queries, retrieving and reading articles, suggesting improvements and elaborating on details to fill reasoning gaps.
- **Coherence, practicality and feasibility improvements.** The agent aims to address issues and creates more coherent hypotheses, potentially rectifying underlying problems with invalid initial assumptions. The agent also refines the hypotheses to make them more practical and feasible. Supplementary Note 9.4 provides an example of the feasibility improvement prompt.
- **Inspiration from existing hypotheses.** The agent additionally creates new hypotheses inspired by single or multiple top-ranked hypotheses.
- **Combination.** The agent also attempts to directly combine the best aspects of several top-ranking hypotheses to create new hypotheses.
- **Simplification.** The agent simplifies hypotheses for easier verification and testing.
- **Out-of-box thinking.** The agent also explores out-of-the-box ideas by moving away from a subset of hypotheses and generating divergent ones. Supplementary Note 9.4 provides an example prompt for this.

The Evolution agent generates new hypotheses; it does not modify or replace existing ones. This strategy protects the quality of top-ranked hypotheses from flawed improvements, as each new hypothesis must also compete in the tournament. The evolution of research hypotheses and proposals also allows Co-Scientist to iteratively combine different improvement techniques and gradually improve the quality of the results.

Meta-review agent

The Meta-review agent serves a crucial role in Co-Scientist's feedback loop, enabling self-improvement in scientific thinking and reasoning. This agent operates on the tournament state and summarizes common patterns identified in reviews and scientific debates in the tournament matches into a meta-review critique.

By synthesizing insights from all reviews, the Meta-review provides valuable feedback to the Reflection agent, leading to more thorough and reliable future reviews. This helps to prevent oversight of critical details. Consider the illustrative example of a identifying a repurposing drug candidate for ALS as a research goal: while only 90% of individual reviews might correctly identify a blood-brain barrier permeability issue in a proposed candidate, the meta-review ensures that all future reviews by the Reflection agent definitively address this crucial factor. Hypothesis and research proposal generation is also enhanced by the meta-review's identification of recurring issues. While the Generation agent uses this feedback selectively to avoid over-fitting to these review critiques, it helps to prevent the recurrence of common issues.

Supplementary Note 9.5 provides an example prompt for the meta-review. In Supplementary Note 10.10, we showcase an example of the summarized meta-review critique generated for the reviews of the previously introduced ALS mechanism hypotheses.

Research overview generation

At the end of the Co-Scientist computation, the Meta-review agent synthesizes top-ranked hypotheses into a research overview, providing a roadmap for future research. This overview outlines potential research areas and directions relevant to the research goal, justifying their importance and suggesting specific experiments within each. Each area includes illustrative example topics. The research overview also serves as an additional input to the Generation agent in subsequent iterations. The research overview serves to effectively map the boundary of current knowledge relevant to the research goal in Co-Scientist and helps highlight future areas of exploration. In Supplementary Note 10.11, we show an example of a research overview for the ALS mechanism research goal. The Meta-review agent can further format these overviews using constrained decoding techniques³⁶ to adhere to common research publication and grant formats (for example, US National Institutes of Health Specific Aims Page format). We demonstrate the effectiveness of this in subsequent sections.

Research contacts identification

The Meta-review agent also uses previous literature reviews to suggest qualified domain experts for research hypothesis and proposal review, including the reasoning behind each suggestion. These potential contacts are summarized in the research overview, providing researchers with additional perspectives and potential avenues for collaborations. An example research contact (with the researcher name redacted) is shown in Supplementary Note 10.12.

Expert-in-the-loop interactions with Co-Scientist

Co-Scientist empowers scientists to actively steer and guide the system through an expert-in-the-loop design (Fig. 1a,b). Scientists can interact with the system in several ways. The typical interaction between Co-Scientist and a human follows a structured process:

- **Research goal definition:** the process begins with a scientist defining the high-level research objective. This involves writing a detailed prompt that can include the specific research question, known constraints of the hypotheses solution space, desired attributes of the output and relevant background literature and data. With proper goal definition, scientists can direct Co-Scientist to follow up on specific research directions (for example, restricted to a smaller collection of prior publications). When this research is referenced in the research goal, Co-Scientist can prioritize generation methods that can access and synthesize it.
- **Goal refinement:** the scientist can refine the initial research goal in light of the generated hypotheses and research overview.
- **Providing review:** the scientist can also provide manual reviews of generated hypotheses, which Co-Scientist uses to evaluate and improve the hypotheses and proposals.
- **Providing ideas and hypotheses:** in the user interface, scientists are allowed to contribute their own hypotheses and proposals for inclusion in the tournament, where they are ranked alongside and can be combined with system-generated hypotheses and proposals.
- **Final review and selection:** after the Co-Scientist run is complete, the scientist is presented with a ranked list of hypotheses and a synthesized research overview from the Meta-review agent. The expert then invests time in reviewing the top-ranked proposals to select the most promising candidates for further experimental validation.

This workflow empowers scientists to guide Co-Scientist at critical junctures. As illustrative examples, we quantified the human time investment for our main validation studies. For the AML drug-repurposing study, the initial prompt, defining the goal to find novel combination therapies, required less than 1 h of an expert clinician's time. After the system's complex run, the final review and selection of promising candidates for in vitro testing took roughly 3 h. Similarly, the fibrosis

target discovery and AMR mechanism generation tasks each required comparable, similar time investments from experts for setup and final review. The scientists and experts featured in our validations have noted that Co-Scientist accomplishes work that would otherwise require days and even weeks of the scientists' time.

Tool use in Co-Scientist

Co-Scientist leverages various tools during the generation, review and improvement of hypotheses. Web search and retrieval are primary tools, important for grounded, up-to-date hypotheses. For research goals that explore a constrained space of possibilities (for example, all known cell receptors of a specific type or all FDA-approved drugs), Co-Scientist agents use domain-specific tools, such as open databases, to constrain searches and generate hypotheses. Co-Scientist can also index and search a private repository of publications and experimental data specified by the scientist. Finally, the system can use and incorporate feedback from specialized AI models, such as AlphaFold. We demonstrate this qualitatively with a protein design example in the Supplementary Note 11 and Supplementary Fig. 9.

Ablation analysis

To validate the contributions of Co-Scientist's core components and agents, we performed a series of quantitative ablation studies (Supplementary Note 3 and Supplementary Figs. 2–6). These analyses revealed that our architectural choices provide tangible benefits to performance and robustness of the overall system.

Specifically, we quantified the value of our multi-faceted Generation agent, showing that a diverse set of generation strategies contributes to the creation of correct hypotheses across different benchmarks.

Critically, the Reflection agent's ability to assess novelty was shown to be dependent on its integration with an external search tool. On a dataset of published ideas that should be rated as non-novel, the agent without search incorrectly assigned a high auto-evaluation novelty score of 6.14 (out of 10), while the agent with search correctly assigned a low novelty score of 2.38 (out of 10). The search tool also enhanced review accuracy, raising the average auto-evaluation correctness score from 7.4 to 8.46 (out of 10) for these known-correct ideas and increasing the area under the curve (AUC) on the GPQA benchmark from 0.643 to 0.651 in our run using Gemini 2.0 Flash as the base foundational LLM for the agents in the system.

Furthermore, we demonstrated that using the scientific debate prompt rather than simple comparison prompt for the Ranking agent can significantly enhance ranking accuracy for high-quality hypotheses and, critically, reduces the positional bias.

Our analysis of the Evolution agent confirms that its iterative refinement process is crucial, boosting precision on GPQA from 70.9% to 75.4%, and increasing the average research hypothesis quality score (auto-evaluation) on the constructed dataset from 4.7 to 5.6.

The effectiveness of our Proximity agent was also validated, as its semantic similarity scores showed a strong correlation with the actual quality score differences between hypotheses, ensuring a comprehensive exploration of the problem space.

Finally, we found that Meta-review agent measurably improves the quality of correctness reviews, increasing AUC for predicting a correct solution from 0.521 to 0.597 in our constructed dataset, and 0.629 to 0.634 in the GPQA diamond dataset.

These results provide important quantitative evidence and support that our multi-agent architecture and specialized prompting strategies are critical design choices that lead to more accurate, robust and reliable scientific hypothesis generation and evaluation.

Statistics and reproducibility

No statistical methods were used to predetermine sample sizes. For computational evaluations, sample sizes ($n = 203$, $n = 15$ and $n = 11$ research goals) were chosen to ensure robust statistical averaging and broad

representation across diverse scientific domains. For in vitro validations, five distinct AML cell lines were tested in independent biological triplicates ($n = 3$). This sample size was not predetermined by statistical methods but was chosen on the basis of widely accepted standard practices for preliminary in vitro dose-response viability screening. Given the large effect sizes typical of such preliminary pharmacological assays, three independent biological replicates provide the necessary degrees of freedom to calculate s.d., assess assay consistency and robustly fit nonlinear regression curves for IC₅₀ estimation. All attempts at replication in both computational and in vitro were successful. The human expert evaluation of the LLM-generated outputs was explicitly blinded, ensuring that independent domain experts were completely unaware of which model generated the hypotheses they were scoring. Blinding was not applicable to the in vitro cell viability screening assays because these experiments involve standardized, automated multimode microplate reader readouts. All cell lines were authenticated by their respective providers and were confirmed negative for mycoplasma contamination.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Except for the three real-world validation tasks (drug repurposing for AML, novel-target discovery for liver fibrosis, mechanism explanation of gene transfer evolution), the remaining datasets used for development, benchmarking and evaluation of the systems are open source or otherwise accessible publicly with permissions. Specifically, the GPQA diamond dataset is publicly available at Hugging Face (<https://huggingface.co/datasets/Idavidrein/gpqa>). The Cancer Dependency Map (DepMap) Q2 2024 data used for computational sanity checks are publicly available at the DepMap portal (<https://depmap.org/portal/>). The curated drug targets dataset from the Open Targets Platform is available online (<https://platform.opentargets.org/downloads>).

Code availability

The full source code for the Co-Scientist system is not publicly available. Owing to the deep integration of the Co-Scientist multi-agent framework with proprietary internal infrastructure, the immense computational resources required for massive test-time scaling and the safety implications of unmonitored autonomous agentic use of such capable AI systems, we are unable to publicly release the full source code or provide broad access immediately. Instead, to enable and accelerate research on important scientific problems, we are initiating an experimental access program. We request scientists interested in solving important scientific problems to reach out and we will provision access subject to computational resources. In due course, we expect to provide broader access to thousands of scientists across the scientific community through Google APIs and bespoke interfaces over time. To further aid transparency and reproducibility, we have provided comprehensive pseudocode detailing the multi-agent orchestration and tournament logic (Supplementary Note 8), alongside the exact system prompts used (Supplementary Note 9). The foundational LLM model, Gemini, used in Co-Scientist is publicly available through APIs, Google AI studio, the Gemini App and other surfaces. Furthermore, important technical details on Gemini have been described extensively in corresponding technical reports^{8,14,37}. The system, algorithms and analysis scripts were developed using Python (v.3.11.7). Data handling and visualization were performed using pandas (v.2.1.4), numpy (v.1.26.4), seaborn (v.0.12.2) and matplotlib (v.3.8.0). For in vitro validations, dose-response curve fitting and IC₅₀ estimation were performed using GraphPad Prism (v.10.6.0, GraphPad Software). Dose-effect data from single-agent and combination treatments were analysed using Julius AI statistical software (accessed November 2025).

28. Jones, N. OpenAI's 'deep research' tool: is it useful for scientists? *Nature* <https://doi.org/10.1038/d41586-025-00377-9> (2025).
29. Megat, S. et al. Integrative genetic analysis illuminates ALS heritability and identifies risk genes. *Nat. Commun.* **14**, 342 (2023).
30. Elo, A. E. *The Rating of Chess Players, Past and Present* (Ishi Press, 2008).
31. LeCun, Y. A theoretical framework for back-propagation. In *Proc. 1988 Connectionist Models Summer School* Vol. 1 (eds Touretzky, D. et al.) 21–28 (Morgan Kaufmann, 1988).
32. Hamilton, A. H., Kalenkova, A. & Roughan, M. The impact of intransitivity on the Elo rating system. *PLoS One* **20**, e0338261 (2025).
33. Coulom, R. Computing 'Elo ratings' of move patterns in the game of go. *ICGA J.* **30**, 198–208 (2007).
34. Kovalchik, S. Extension of the Elo rating system to margin of victory. *Int. J. Forecast.* **36**, 1329–1341 (2020).
35. Khan, A. et al. Debating with more persuasive LLMs leads to more truthful answers. In *Proc. 41st International Conference on Machine Learning* Vol. 235 (eds Salakhutdinov, R. et al.) 23662–23733 (PMLR, 2024).
36. Post, M. & Vilar, D. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1 (Long Papers) 1314–1324 (Association for Computational Linguistics, 2018).
37. Gemini Team, Google. Gemini: a family of highly capable multimodal models. Preprint at <https://arxiv.org/abs/2312.11805> (2023).

Acknowledgements We thank our teammates S. Venugopalan, J. Platt, E. Brand and Y. Liu for detailed technical feedback on the manuscript; J. T. Rostoele, C. Chmielowska, J. B. Patkowski, J. Inchai, W. Liu and W. Ren for providing expert feedback on the AI system introduced in this work; the laboratory of R. Majeti from Stanford University for providing the AML cell lines used in this work; R. Raman, R. Flynn, C. Hempstead, A. Darzi, O. Abudayyeh, J. Gootenberg, N. Fishman, J. Lequyer, D. Leesman, R. Solanki, D. Gong and A. Sadagopan for feedback on different aspects of the AI system and the work; M. Abdelrahim, E. Burns, P. Prasad and H. Mai for clinical expertise and expert evaluation; H. Gowda, C. Balagtas and the Signios Biosciences team for support with the wet-laboratory experiments; our teammates A.-C. Rivers, S. Basu, S. Nowozin, T. Wagner, N. Latysheva, N. Kerem, Y. Carmel, H. H. Harriou, L. Tamulevičius, I. Grublyte, T. Applebaum, M. Shah, N. Stroppa, M. Ciorobea, J. Oesighaus, D. Ballesteros, L. Vazić, A. Trostanetski, B. Byrne, B. James, J. Barrios, I. Lee, T. Rakket and B. Panichprecha for technical support; C. Taylor, J. Walker, J. Mason, H. Gladman, S. Singh, I. Bowers, A. Cupani, F. Pietra, U. Okereke, A. Senoner, A. Fernandez Lara, J. Bawagan, D. Breen, S. Lee, D.-P. Barley, K. Tong, Q. Green, E. Rolland, T. Goddu, R. Parikh, S. Kou, R. Sico, A. Ferber, C. Kozłowski, A. Lentz, K. K. Walker, R. Ruparel, J. Sturgeon, L. Verburgh, K. Seager, L. Karlsson, D. Vijaykumar, L. Winer, E.-A. Graham, T. Milner, M. K. Blake, E. Radhansson, I. Ghosh, J. Nayar, B. Cappy, C. Grade, A. Jones, L. Vardoulakis, L. Dorfman, A. Chakraborty, D. Williams-Falokun, M. Shiels, K. Pamarthy, S. Brown, A. Song, C. Wright, E. Zadina, G. Hong, M. Klein, S. Leung, J. Gaiser, R. Green, V. Johnston, M. de Andres-Clavera, A. Singh, A. Dondzik, T. Beyer, D. Russell, V. Vuskovic, G. Jennings, K. Schaefer, M. Felix, R. Teo, S. Giardino, K. Paterson, A. Laydon, V. Langston, M. Voges, A. Otter, S. Pickering, P. Kunkle, M. H. Taege, A. Marsh, A. Mould, J. Fenster, F. Garcia, Z. Wang, P. Beak, J. Stevenson, A. Brik and S. S. Mahdavi for support and guidance during the course of this project; and M. Brenner, Z. Ghahramani, D. Webster, J. Barral, M. Howell, S. Thomas, K. DeSalvo, J. Freidenfelds, R. Levavi Morad, A. Esлами, A. Koivuniemi, G. Corrado, R. Hansen, A. Berndt, S. Narayanan, C. Mayer, P. Ranganathan, A. Jain, S. Huffman, J. Woodward, J. Jumper, N. Shazeer, O. Vinyals, K. Kavukcuoglu, T. Kurian, J. Dean and S. Pichai for support of this work.

Author contributions J.G., V.N., T.T. and A.K. initiated the project. J.G., W.-H.W., A.D., T.T., Y.G., V.D., E.D.V., B.L., T.R.D.C., J.R.P., G.P., A. Pawlosky, A.K. and V.N. contributed to the conception of the study and study design. J.G., W.-H.W., A.D., P.S., A.M., G.G., F.W., A.O., D.P., J.B., D.Z., I.R., E.V., F.H., L.R., M. Boia, I.B., B.F., M. Bellaiche, T.S., J.F., J.R., B.G., A.V., D.H., Y.X., P.K., A. Pawlosky, A.K. and V.N. contributed to system design. J.G., W.-H.W., A.D., T.T., O.B., Y.G., V.D., E.D.V., B.L., T.R.D.C., J.R.P., G.P., P.K., A. Pawlosky, A.K. and V.N. contributed to acquisition of the data. J.G., W.-H.W., A.D., T.T., P.S., A.M., F.W., G.G., A. Palepu, D.P., F.Z., Y.G., V.D., E.D.V., B.L., T.R.D.C., J.R.P., G.P., A. Pawlosky, A.K. and V.N. contributed to analysis and interpretation of the data. J.G., K.R., J.B., A.C., K.K., K.C., A.H., B.G., A.V., Y.M., J.M., D.H., Y.X., P.K., A. Pawlosky, A.K. and V.N. provided strategic guidance. Y.X. led product discovery and development. J.G., W.-H.W., A.D., T.T., K.K., O.B., Y.X., A. Pawlosky, A.K. and V.N. contributed to paper organization and team logistics. J.G., J.R., N.T., O.B. and V.N. contributed to safety evaluation. J.G., W.-H.W., T.T., P.S., A. Palepu, R.T., K.S., N.T., J.R., V.D., E.D.V., B.L., J.M., P.K., A. Pawlosky, A.K. and V.N. contributed to drafting and revising the manuscript. W.-H.W. led manuscript preparation.

Funding This work was supported by US National Institutes of Health awards (1R01DC021133 and 1R24OD035408) to G.P.

Competing interests This study was funded by Alphabet and/or a subsidiary thereof. Some authors (J.G., W.-H.W., AD, T.T., P.S., A.M., G.G., F.W., A.O., D.P., A. Palepu, K.R., R.T., K.S., F.Z., A.C., K.K., N.T., D.Z., I.R., E.V., F.H., L.R., M. Boia, I.B., B.F., M. Bellaiche, T.S., J.F., J.R., O.B., K.C., A.H., B.G., A.V., Y.M., J.M., D.H., P.K., Y.X., A. Pawlosky, A.K. and V.N.) are employees of Alphabet and may own stock as part of the standard compensation package. E.D.V. and B.L. are employees of Sequome. B.L. is the founder of Dendra Therapeutics.

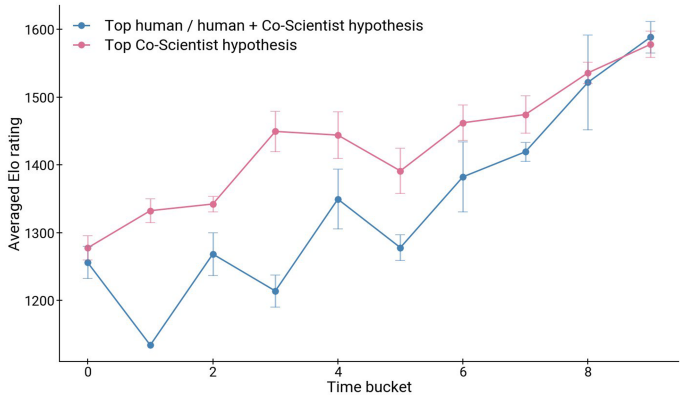
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-026-10644-y>.

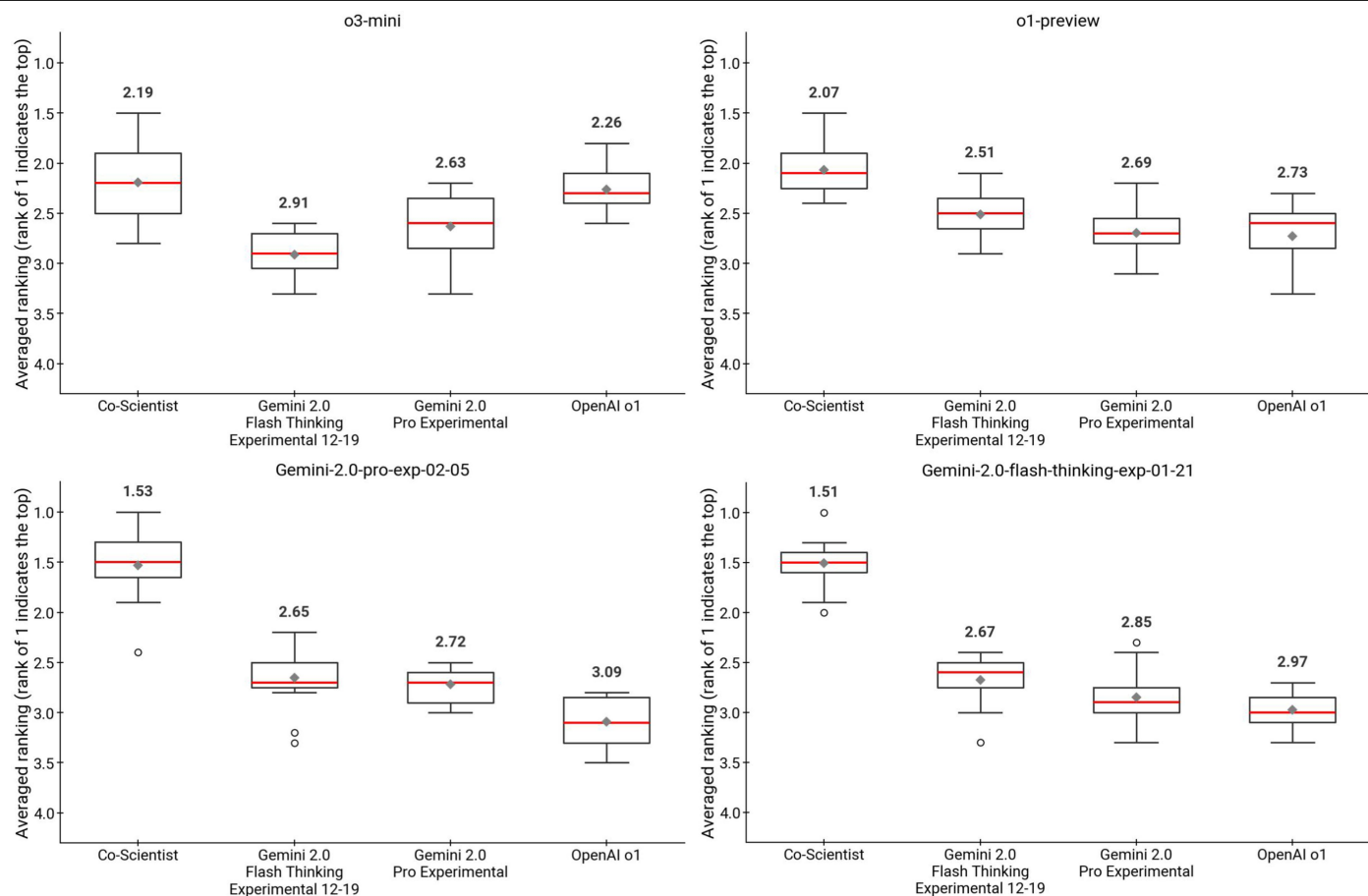
Correspondence and requests for materials should be addressed to Juraj Gottweis, Wei-Hung Weng, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam or Vivek Natarajan.

Peer review information Nature thanks Olivier Elemento, Linfeng Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

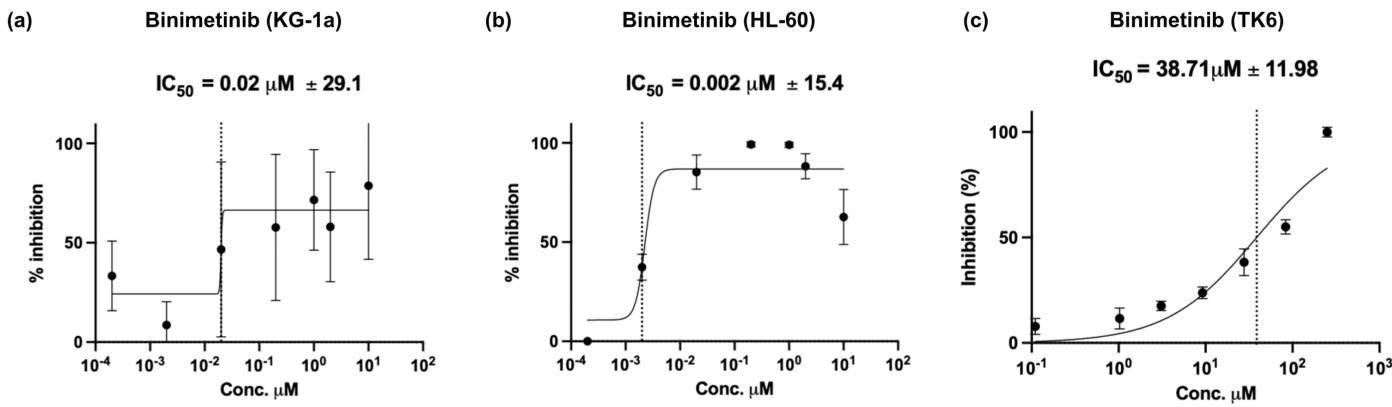


Extended Data Fig. 1 | AI-augmented expertise with Co-Scientist through Elo-based auto-evaluation. Through its self-improvement process, Co-Scientist refines and enhances expert ‘best guess’ solutions over time, as measured by the Elo rating on a subset of 15 curated research goals. It is important to note that the Elo metric is auto-evaluated and not based on independent ground truth. The error bar indicates the corresponding standard error of the mean (s.e.m.). Data are presented as mean $\pm$ s.e.m. for 15 independent expert-curated research goals.



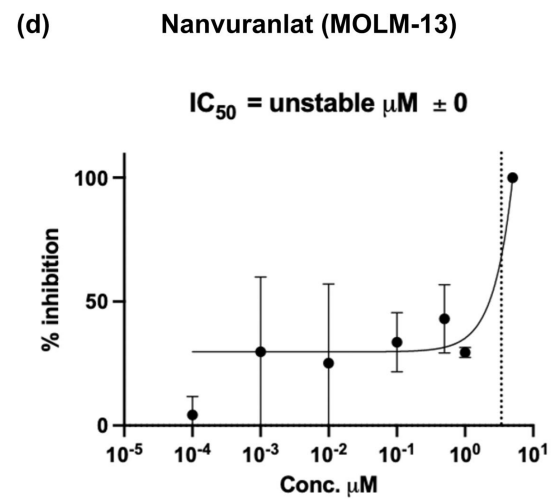
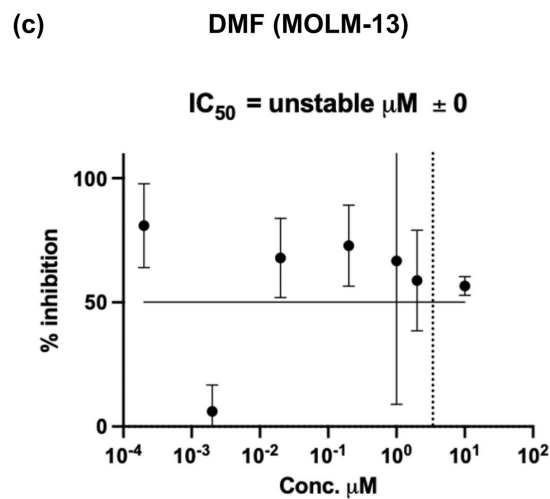
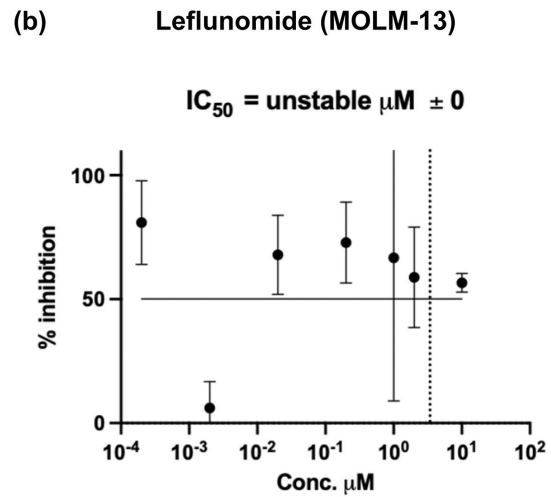
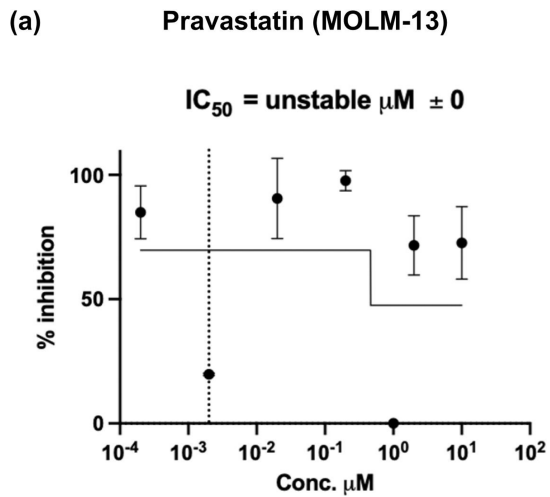
Extended Data Fig. 2 | LLM preference ranking auto-evaluation of Co-Scientist and other baselines. Averaged preference ranking of results across 15 expert curated research goals generated by Co-Scientist, Gemini 2.0 Flash Thinking Experimental 12-19, Gemini 2.0 Pro Experimental, and OpenAI o1, using four different LLM evaluators: OpenAI o3-mini-2025-01-31 (upper left), OpenAI o1-preview-2024-09-12 (upper right), Gemini 2.0 Pro Experimental (lower left), and Gemini 2.0 Flash Thinking Experimental 01-21 (lower right).

Lower numbers indicate better rankings. In each box plot, the central line represents the median ranking, the top and bottom edges of the box indicate the 25th and 75th percentiles respectively, the whiskers extend to the most extreme data points within 1.5 times the interquartile range, and the grey diamond marks the mean. Each box plot represents averaged evaluation scores for the 15 independent expert-curated research goals.



Extended Data Fig. 3 | Dose response curves of drug repurposing candidate Binimetinib in other cell lines. Binimetinib demonstrates activity inhibiting cell viability in KG-1a (a), HL-60 (b) and TK6 (c) cell lines. x axis is the drug concentration (μM), and y axis is the percentage of growth inhibition. Binimetinib's target is related to the RAS-RAF-MEK-ERK pathway which is not

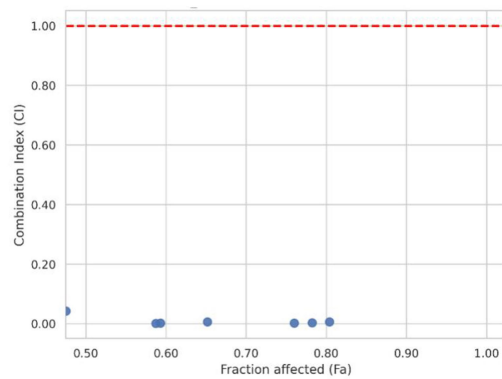
generally expected to be overactive or essential in the TK6 cell line compared to AML cells. Thus, the significantly higher IC_{50} in TK6 compared to three AML cell lines (MOLM-13, KG-1a, HL-60), is consistent with its intended mechanism of action. Data are presented as mean $\pm$ s.d. of $n = 3$ biologically independent experiments.



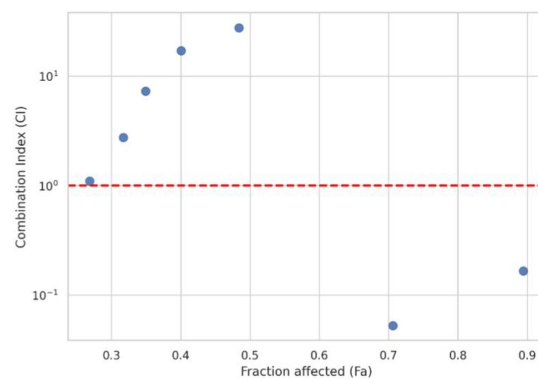
Extended Data Fig. 4 | Dose response curves of the drug repurposing candidates with little to no effect on MOLM-13. Of the expert-selected drug repurposing candidates, Pravastatin (a) and dimethyl fumarate (DMF) (c) showed little to no effect on the MOLM-13 cell line across the concentrations tested. Of the novel drug repurposing candidates, Leflunomide (b) and

Nanvuranlat (d) showed little to no effect on the MOLM-13 cell line across the concentrations tested. x axis is the drug concentration (μM), and y axis is the percentage of growth inhibition. Data are presented as mean $\pm$ s.d. of $n = 3$ biologically independent experiments.

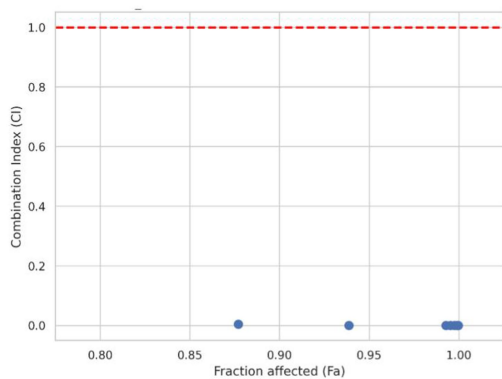
(a) JNJ-64619178 + SNDX-5613 (MOLM-13)



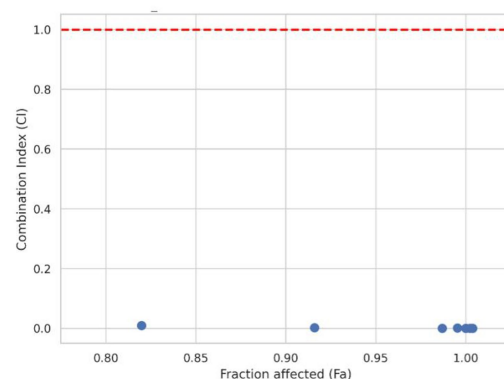
JNJ-64619178 + SNDX-5613 (KG-1a)



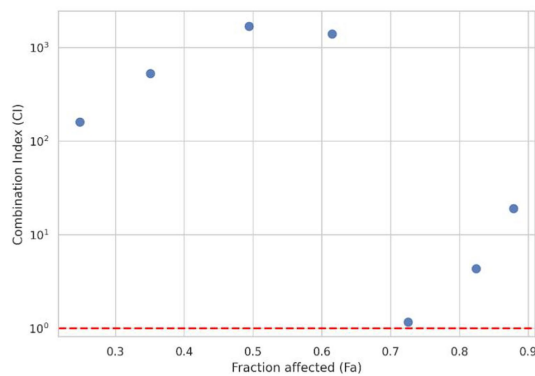
(b) Palbociclib + Selinexor (MOLM-13)



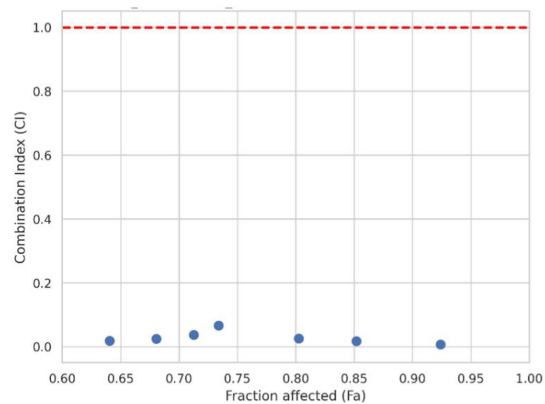
Palbociclib + Selinexor (KG-1a)



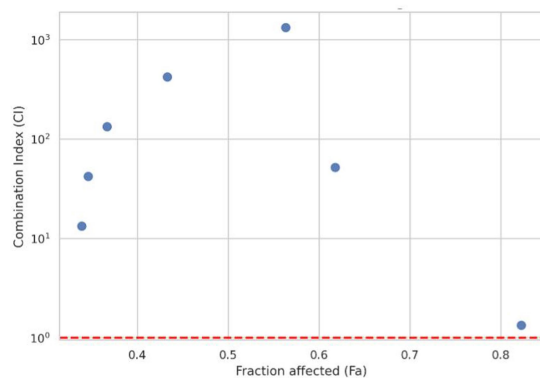
(c) Venetoclax + Pinometastat (MOLM-13)



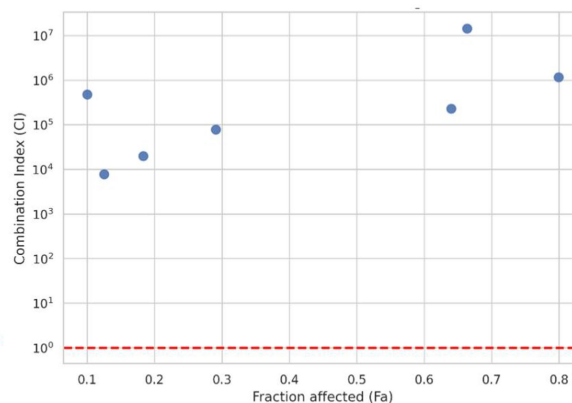
Venetoclax + Pinometastat (KG-1a)



(d) CB-839 + Sulfasalazine (MOLM-13)



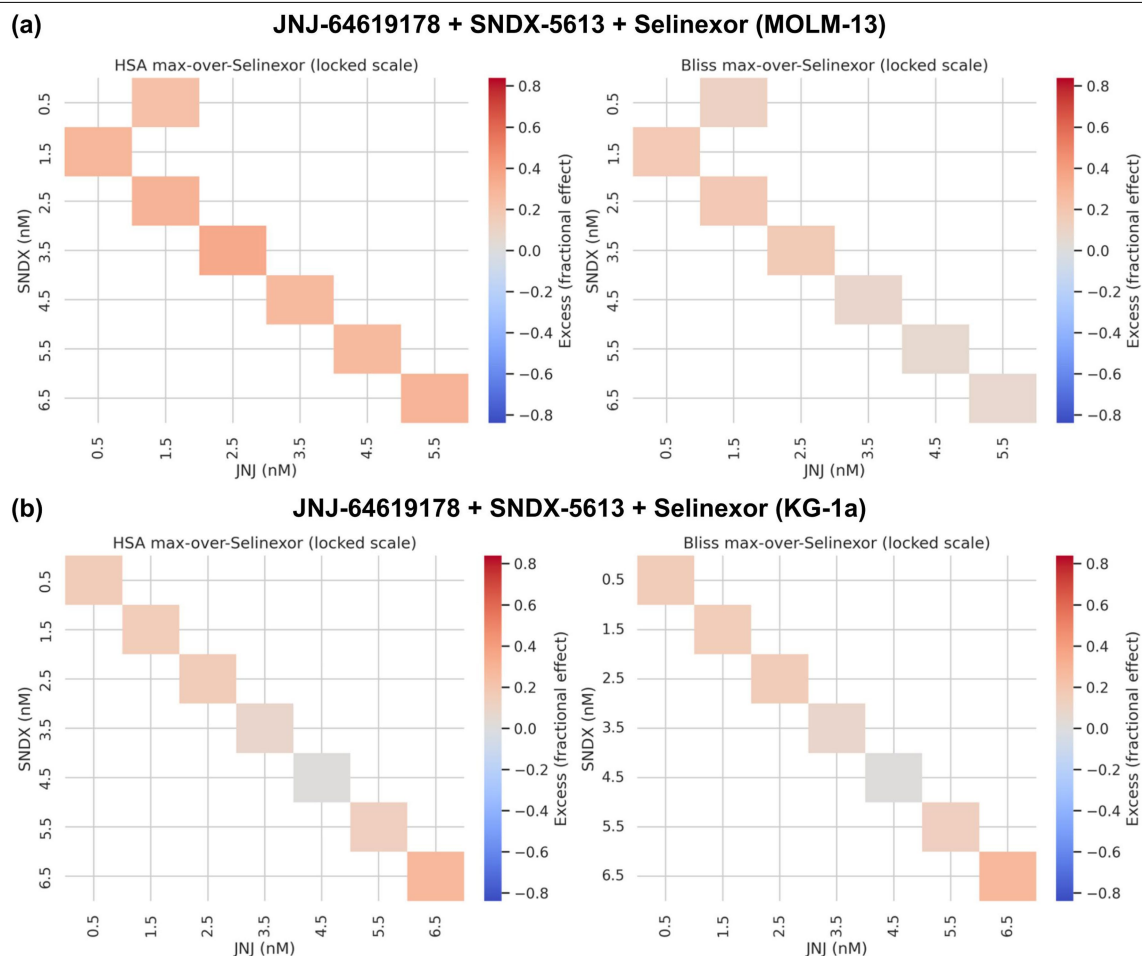
CB-839 + Sulfasalazine (KG-1a)



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Comprehensive synergy analysis of Co-Scientist nominated dual drug combinations in AML cell lines. Building upon the representative example in Fig. 4a,b, this figure details the pharmacological interactions of additional AI-predicted doublets evaluated in MOLM-13 (left panels) and KG-1a (right panels) cell lines. Synergy was quantified using the Chou-Talalay method based on fixed-ratio dose-response data. Each panel presents a Fraction affected versus Combination Index (Fa-CI) plot.

The horizontal red dashed line indicates a strictly additive effect (CI = 1.0). Data points below the line (CI < 1.0) indicate synergistic interactions, while points above the line (CI > 1.0) indicate antagonism. The results highlight that while certain combinations (e.g., Palbociclib + Selinexor **(b)**) exhibit consistent synergy across different genetic backgrounds, others display context-dependent interaction profiles strongly influenced by the specific cell line's mutational status.



Extended Data Fig. 6 | Excess fractional effect heatmaps for Co-Scientist predicted triple drug combinations. Complementing Fig. 4c,d, this figure provides the interaction matrices for the other evaluated triplet regimen (JNJ-64619178 + SNDX-5613 + Selinexor) across AML cell lines MOLM-13 (a) and KG-1a (b). Synergy and antagonism are quantified across a comprehensive grid of drug concentrations (nM) utilizing both the Highest Single Agent (HSA) and Bliss independence models. The heatmaps display the fractional excess effect:

the difference between the observed combined empirical effect and the predicted additive effect. Positive excess effects are visualized in red (synergy), zero excess in white (additivity), and negative excess in blue (antagonism). To compress the three-dimensional dose space into a two-dimensional matrix, the plots display the maximum positive interaction across the third drug's dose range.

Extended Data Table 1 | IC₅₀ of the drugs used in the drug combination experiments in AML cell lines, MOLM-13 and KG-1a

Drug	IC ₅₀ (μM)	
	MOLM-13	KG-1a
MSA-2	No response	No response
Sulfasalazine	No response	No response
CB-839	16.98	8.79
JNJ-64619178 (Onametastat)	0.0046	0.024
Pinometastat	No response	No response
Selinexor	0.039	0.055
JQ1	0.011	0.12
Venetoclax	0.009	0.006
Olaparib	2.12	7.1
SNDX	0.010	No response
Palbociclib	0.026	0.07

The table presents the half-maximal inhibitory concentrations (IC₅₀, in μM) for 11 individual drugs evaluated in two acute myeloid leukaemia (AML) cell lines, MOLM-13 and KG-1a. These single-agent efficacies were determined to establish the baseline dose-response behaviours required for the subsequent drug combination synergy analyses. Cell viability was assessed following 72 to 96 h of drug exposure. IC₅₀ values were estimated using non-linear regression curve fitting. ‘No response’ indicates that the drug failed to achieve 50% growth inhibition within the tested concentration range. All experiments were performed in biologically independent triplicates (n=3).

Extended Data Table 2 | Summary of the drug combination synergy experiments

Drug combination	MOLM-13	KG-1a
JNJ-64619178 + Selinexor	Synergistic	Mixed response
Palbociclib + Selinexor	Synergistic	Synergistic
JNJ-64619178 + SNDX-5613	Synergistic	Mixed response
CB-839 + Sulfasalazine	Not synergistic	Not synergistic
Venetoclax + Pinometostat	Mixed response	Synergistic
JNJ-64619178 + Selinexor + SNDX-5613	Synergistic	Synergistic
JQ1 + Olaparib + MSA-2	Synergistic	Not synergistic

The table categorizes the overall in vitro interaction profiles of seven Co-Scientist nominated drug combinations, including doublet and triplet combinations, evaluated in two acute myeloid leukaemia (AML) cell lines, MOLM-13 and KG-1a. Pharmacological interactions were quantitatively assessed using the Chou-Talalay combination index (CI) method for doublets, and Highest Single Agent (HSA) and Bliss independence models for triplets. The interaction outcomes are classified into three categories, ‘Synergistic’ (exhibiting $CI < 1$ or positive excess fractional effect), ‘Not synergistic’ (additive or antagonistic effects), and ‘Mixed response’ (exhibiting context-dependent effects, transitioning between synergy and antagonism depending on the dose or effect level). All experiments were performed in biologically independent triplicates ($n=3$).

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	The system, algorithms and scripts were implemented using Python 3.11.7. Scientific literature and data were collected via web search APIs and internal data pipelines compatible with the Gemini 2.0 model infrastructure. Data collection for in vitro cell viability was performed using a multimode microplate reader.
Data analysis	The full source code for the Co-Scientist system is not publicly available. The foundational LLM model, Gemini, used in Co-Scientist is publicly available via APIs, Google AI studio, the Gemini App and other surfaces. The analysis codes were developed using Python (version 3.11.7). Data handling and visualization were performed using pandas (version 2.1.4), numpy (version 1.26.4), seaborn (version 0.12.2), and matplotlib (version 3.8.0). For in vitro validations, dose-response curve fitting and IC50 estimation were performed using GraphPad Prism (version 10.6.0, GraphPad Software). Synergy analysis and Combination Index calculations were performed using Julius AI statistical software (accessed November 2025).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Except for the three real-world validation tasks (drug repurposing for AML, novel target discovery for liver fibrosis, mechanism explanation of gene transfer evolution), the remaining datasets used for development, benchmarking and evaluation of the AI systems are open source or otherwise accessible publicly with permissions. Specifically, the GPQA diamond dataset is publicly available at Hugging Face (<https://huggingface.co/datasets/ldavidrein/gpqa>). The Cancer Dependency Map (DepMap) Q2 2024 data used for computational sanity checks is publicly available at the DepMap portal (<https://depmap.org/portal/>). The curated drug targets dataset from the Open Targets Platform is available at <https://platform.opentargets.org/downloads>.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Reporting on race, ethnicity, or other socially relevant groupings

Population characteristics

Recruitment

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample sizes. For the computational evaluations, the sample size of 203 diverse research goals, 15 expert-curated goals, and 11 goals for LLM-as-a-judge preference were chosen to ensure a broad representation of scientific domains, robust statistical averaging across varying complexity, and to mitigate evaluator bias. For the in vitro validations, five distinct AML cell lines (MOLM-13, KG-1a, HL-60, NOMO-1, and TK6) were tested to model the molecular and genetic heterogeneity of the disease. Experiments were performed in biological triplicates (n=3). This sample size was not predetermined by statistical methods but was chosen based on widely accepted standard practices for preliminary in vitro dose-response viability screening. Given the large effect sizes typical of such preliminary pharmacological assays, three independent biological replicates provide the necessary degrees of freedom to calculate standard deviation, assess assay consistency, and robustly fit non-linear regression curves for IC50 estimation.
Data exclusions	No data were excluded from the computational or in vitro laboratory analyses.
Replication	All in vitro dose-response (single-agent screening) and cell viability assays (fixed-ratio combination treatments) were performed in independent biological triplicates (n=3). For computational hypothesis generation and LLM-as-a-judge evaluations, the system generated responses or evaluations independently across multiple discrete runs. All attempts at replication (both in vitro and computational) were successful. The resulting in vitro data were presented as the mean $\pm$ standard deviation of these triplicates.
Randomization	Randomization was not relevant for the in vitro experiments, as these involved standardized dose-response testing where specified drug concentrations were deterministically applied to cell cultures. For the human evaluation of the Co-Scientist and other LLMs, outputs were assessed by experts, the model outputs were presented randomly.
Blinding	The human expert evaluation of the LLM-generated outputs was explicitly blinded, ensuring that independent domain experts were completely unaware of which model generated the hypotheses they were scoring. Blinding was not applicable to the in vitro cell viability screening assays because these experiments involve standardized, automated readouts (absorbance at 490 nm using a multimode microplate

reader) that are objective and not subject to investigator interpretation bias. Furthermore, knowledge of the drug identity is required by the experimenter to prepare the appropriate molar concentration ranges and dilution series for accurate IC50 determination.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☐	☒ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☒	☐ Public health
☒	☐ National security
☒	☐ Crops and/or livestock
☒	☐ Ecosystems
☒	☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
☒	☐ Demonstrate how to render a vaccine ineffective
☒	☐ Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐ Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐ Increase transmissibility of a pathogen
☒	☐ Alter the host range of a pathogen
☒	☐ Enable evasion of diagnostic/detection modalities
☒	☐ Enable the weaponization of a biological agent or toxin
☒	☐ Any other potentially harmful combination of experiments and agents

Seed stocks	Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.
Novel plant genotypes	Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.
Authentication	Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.