Towards an AI co-scientist

Juraj Gottweis[*, ‡, 1], Wei-Hung Weng[*, ‡, 2], Alexander Daryin[*,1], Tao Tu[*,3],
Anil Palepu[2], Petar Sirkovic[1], Artiom Myaskovsky[1], Felix Weissenberger[1],
Keran Rong[3], Ryutaro Tanno[3], Khaled Saab[3], Dan Popovici[2], Jacob Blum[7], Fan Zhang[2],
Katherine Chou[2], Avinatan Hassidim[2], Burak Gokturk[1],
Amin Vahdat[1], Pushmeet Kohli[3], Yossi Matias[2],
Andrew Carroll[2], Kavita Kulkarni[2], Nenad Tomasev[3], Yuan Guan[7],
Vikram Dhillon[4], Eeshit Dhaval Vaishnav[5], Byron Lee[5],
Tiago R D Costa[6], José R Penadés[6], Gary Peltz[7],
Yunhan Xu[3], Annalisa Pawlosky[1, ‡], Alan Karthikesalingam[2, ‡] and Vivek Natarajan[2, ‡]

[1]Google Cloud AI Research, [2]Google Research, [3]Google DeepMind,
[4]Houston Methodist, [5]Sequome,
[6]Fleming Initiative and Imperial College London,
[7]Stanford University School of Medicine

Scientific discovery relies on scientists generating novel hypotheses that undergo rigorous experimental validation. To augment this process, we introduce an AI co-scientist, a multi-agent system built on Gemini 2.0. The AI co-scientist is intended to help uncover new, original knowledge and to formulate demonstrably novel research hypotheses and proposals, building upon prior evidence and aligned to scientist-provided research objectives and guidance. The system's design incorporates a generate, debate, and evolve approach to hypothesis generation, inspired by the scientific method and accelerated by scaling test-time compute. Key contributions include: (1) a multi-agent architecture with an asynchronous task execution framework for flexible compute scaling; (2) a tournament evolution process for self-improving hypotheses generation. Automated evaluations show continued benefits of test-time compute, improving hypothesis quality. While general purpose, we focus development and validation in three biomedical areas: drug repurposing, novel target discovery, and explaining mechanisms of bacterial evolution and anti-microbial resistance. For drug repurposing, the system proposes candidates with promising validation findings, including candidates for acute myeloid leukemia that show tumor inhibition *in vitro* at clinically applicable concentrations. For novel target discovery, the AI co-scientist proposed new epigenetic targets for liver fibrosis, validated by anti-fibrotic activity and liver cell regeneration in human hepatic organoids. Finally, the AI co-scientist recapitulated unpublished experimental results via a parallel in silico discovery of a novel gene transfer mechanism in bacterial evolution. These results, detailed in separate, co-timed reports, demonstrate the potential to augment biomedical and scientific discovery and usher an era of AI empowered scientists.

1 Introduction

Human ingenuity and creativity propel the advancement of fundamental research in science and medicine. However, researchers, particularly in biomedicine, are faced with a breadth and depth conundrum. The complexity of biomedical topics require increasingly deep and specific subject matter expertise, while leaps in insight may still arise from broad knowledge bridging across disciplines. With the rapid rise in scientific publications and the availability of numerous technologies for specialized high-throughput assays, mastery of both discipline-specific depth and trans-disciplinary insight can be challenging.

Despite these challenges, many modern breakthroughs have emerged from trans-disciplinary endeavours. Emmanuelle Charpentier and Jennifer Doudna won the 2020 Nobel Prize in Chemistry for their work on CRISPR [1], which combined techniques and strategies ranging from microbiology to genetics to molecular biology. These benefits of synergy have also been seen beyond experimental biomedicine in numerous other areas of science. Notably, Geoffrey Hinton and John Hopfield combined ideas from physics and neuroscience [2, 3] to develop artificial intelligence (AI) systems, which were awarded the 2024 Nobel Prize in Physics.

∗ *Equal contributions.*
‡ *Corresponding authors: {juro, ckbjimmy, apawlosky, alankarthi, natviv}@google.com*

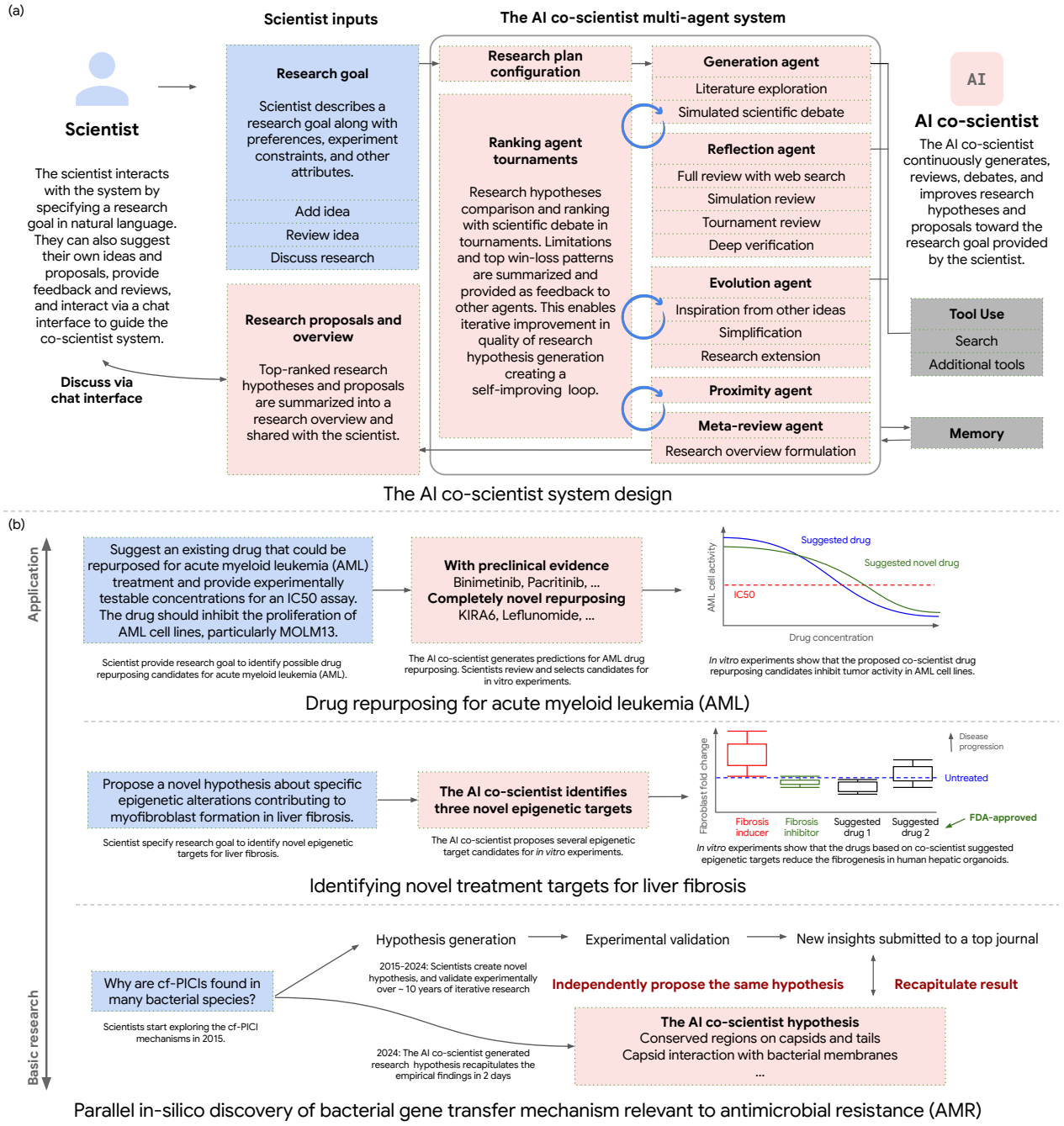

Figure 1 | The AI co-scientist system design and experimental validation summary. (a) Here, we illustrate the different components of the the AI co-scientist multi-agent system, and its interaction paradigm with scientists. Given a research goal in natural language, the co-scientist generates novel research hypotheses and proposals. The system employs specialized agents — Generation, Reflection, Ranking, Evolution, Proximity (which evaluates relatedness), Meta-review (which provides high level analysis) — 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 outputs. The co-scientist leverages tools, including web search and specialized AI models to improve the grounding and quality of generated research hypotheses. Scientists can converse with the co-scientist in natural language to specify research goals, incorporate constraints, provide feedback and suggest new directions for explorations via the designated user interface. (b) We perform end-to-end validation of the co-scientist generated hypotheses in three important topics of biomedicine with varied complexity—suggesting novel drug repurposing candidates for acute myeloid leukemia (AML) (upper panel), discovering novel epigenetic targets for liver fibrosis treatment (middle panel), and recapitulating the discovery of novel mechanism of gene transfer evolution in bacteria key to anti-microbial resistance (lower panel). The co-scientist's hypotheses for these three settings are externally, independently validated by *in vitro* laboratory experiments and detailed in separate preprints co-timed with this work. In the figure, blue denotes expert scientist inputs while red denotes the co-scientist agents or outputs.

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 [4–6], multimodal understanding [6], and agentic behaviors [7] such as the ability to use tools to solve complex tasks over long time horizons. Further, the trends with distillation [8] and inference time compute costs [6, 9] indicate that such intelligent and general AI systems are rapidly becoming more affordable and available. Motivated by the aforementioned unmet needs in the modern discovery process in science and medicine and building on the advancements in frontier AI [10], we develop and introduce an AI co-scientist system.

The co-scientist is designed to act as a helpful assistant and collaborator to scientists and to help accelerate the scientific discovery process. The system is a compound, multi-agent AI system [11] building on Gemini 2.0 and designed to mirror the reasoning process underpinning the scientific method [12]. Given a research goal specified in natural language, the system can search and reason over relevant literature to summarize and synthesize prior work and build on it to propose novel, original research hypotheses and experimental protocols for downstream validations (Figure 1a). The co-scientist provides grounding for its recommendations by citing relevant literature and explaining the reasoning behind its proposals.

This work does not aim to completely automate the scientific process with AI. Instead, the co-scientist is purpose-built for a "scientist-in-the-loop" collaborative paradigm, to help domain experts augment their hypothesis generation process and guide the exploration that follows. Scientists can specify their research goals in simple natural language, including informing the system of desirable attributes for the hypotheses or research proposals it should create and the constraints that the synthesized outputs should satisfy. They can also collaborate and provide feedback in a variety of ways, including directly supplying their own ideas and hypotheses, refining those generated by the system, or using natural language chat to guide the system and ensure alignment with their expertise.

The co-scientist works through a significant scaling of the test-time compute paradigm [13–15] to iteratively reason, evolve, and improve the outputs as it gathers more knowledge and understanding. 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 via the process of finding win and loss patterns, and a hypothesis 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 to provide itself with feedback to iteratively refine its hypotheses and research proposals.

While the co-scientist system is general-purpose and applicable across multiple scientific disciplines, in this study we focus our development and validation of the system to biomedicine. We validate the co-scientist's capability in three impactful areas of biomedicine with varied complexity: (1) drug repurposing, (2) novel treatment targets discovery, and (3) new mechanistic explanations for antimicrobial resistance (Figure 1b).

Drug development is an increasingly time-consuming and expensive process [16] in which new therapeutics require restarting many aspects of the discovery and development process for each indication or disease (roughly 70% of drug approvals are for new drugs). In contrast, drug repurposing——identifying novel therapeutic indications for drugs beyond their original intended use——has emerged as a compelling strategy to address these challenges [17]. Successful examples of repurposing include Humira (adalimumab) and Keytruda (pembrolizumab), both of which have become among the most successful drugs in history. [17]. The process typically involves analyzing molecular signatures, signaling pathways, drug interactions, clinical trial results, adverse event reports, and other literature-based information [18], along with off-label use data and, in some cases, patient experiences. However, drug repurposing is limited by several factors: (1) the need for extensive expertise across biomedical, molecular biology, and biochemical systems, (2) the inherent complexity of mammalian biological systems, and (3) the time-intensive nature of traditional computational biology analyses required. We leverage the co-scientist to generate predictions for large-scale drug repurposing, validating the generated predictions using a combination of computational biology, expert clinician feedback, and *in vitro* wet-lab validation approaches. Notably, our system has proposed novel repurposing candidates for acute myeloid leukamia (AML) that inhibit tumor viability at clinically relevant concentrations *in vitro* across multiple AML cell lines.

Unlike drug repurposing, which is a combinatorial search problem through a large but constrained set of

drugs and diseases, identifying novel treatment targets for diseases presents a more significant challenge, traditionally requiring extensive literature review, deep biological understanding, sophisticated hypothesis generation and complex experimental validation strategies. The uncertainty of identifying novel treatment targets is significantly greater than in drug repurposing, as it involves not only repurposing existing compounds but also uncovering entirely new components and mechanisms within biological systems. This target discovery process can be inefficient, potentially leading to suboptimal hypothesis selection and prioritization for *in vitro* and *in vivo* experimentation. Given the high costs and time associated with experimental validation, a more effective approach is needed. We probe the capabilities of the co-scientist to propose, rank, and provide experimental protocols for novel research hypotheses pertaining to target discovery. To demonstrate this capability, we focus on liver fibrosis, a prevalent and serious disease, showcasing the co-scientist's potential to discover novel treatment targets amenable to experimental validation. In particular, the co-scientist has suggested novel epigenetic targets demonstrating significant anti-fibrotic activity in human hepatic organoids.

As a third validation of the capabilities of our system, we focus on generation of hypotheses to explain mechanisms related to gene transfer evolution in bacteria pertaining to antimicrobial resistance (AMR) - mechanisms developed by microbes to circumvent drug applications used to fight infections. This is arguably an even more complex challenge than drug repurposing and target discovery and involves understanding of not only the molecular mechanisms of gene transfer (conjugation, transduction, and transformation) but also the ecological and evolutionary pressures that drive the spread of AMR genes: a system-level problem with many interacting variables. This is also an important healthcare challenge with increasing rates of infections and deaths worldwide [19]. In this validation, researchers instructed the AI co-scientist to explore a topic that had already been subject to novel discovery by their independent research group. Notably, at the time of instructing the AI co-scientist system, the researchers' novel experimental insights had not yet been published or revealed in the public domain. The system was instructed to hypothesize how capsid-forming phage-inducible chromosomal islands (cf-PICIs) exist across multiple bacterial species. The system independently proposed that cf-PICIs interact with diverse phage tails to expand their host range. This *in silico* discovery mirrored the novel and experimentally validated results that expert researchers had already performed, as detailed in the co-timed report [20, 21].

Overall, our key contributions are summarized as follows:

- **Introducing an AI co-scientist.** We develop and introduce an AI co-scientist that goes beyond literature summarization and "deep research" tools to assist scientists in uncovering new knowledge, novel hypothesis generation and experimental planning.
- **Significant scaling of test-time compute paradigm for scientific reasoning.** The co-scientist is built on a Gemini 2.0 multi-agent architecture, utilizing an asynchronous task execution framework. This framework allows 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 the AI co-scientist outperforming other state-of-the-art (SOTA) 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 the co-scientist in important topics in biomedicine.** We present end-to-end validation of novel AI-generated hypotheses through new empirical findings in three distinct and increasingly complex areas of biomedicine: drug repurposing, novel target discovery, and antimicrobial resistance. The AI co-scientist predicts novel repurposing drugs for AML, identifies novel epigenetic treatment targets grounded in preclinical evidence for liver fibrosis, and proposes novel mechanisms for gene transfer in bacterial evolution and antimicrobial resistance. These discoveries from the AI co-scientist have been validated in wet-lab settings and are detailed in separate, co-timed technical reports.

2 Related Works

2.1 Reasoning models and test-time compute scaling

The modern revolution in foundation AI models [22] and large language models (LLMs) has been largely driven by advances in pre-training techniques [23, 24], leading to breakthroughs in models like the GPT and Gemini family [25, 26]. These models, trained on increasingly massive internet-scale and multimodal datasets, have demonstrated impressive abilities in language understanding and generation leading to breakthrough performance in a variety of benchmarks [27, 28]. However, a key area of ongoing development is enhancing their *reasoning* capabilities. This has led to the emergence of "reasoning models" which go beyond simply predicting the next word and instead attempt to mimic human thought processes [29]. One promising direction in this pursuit is the test-time compute paradigm. This approach moves beyond solely relying on the knowledge acquired during pre-training and allocates additional computational resources during inference to enable System-2 style thinking—slower deliberate reasoning to reduce uncertainty and progress optimally towards the goal [30]. This concept emerged with early successes such as AlphaGo [15], which used Monte Carlo Tree Search (MCTS) to explore game states and strategically select moves, and Libratus [14], which employed similar techniques to achieve superhuman performance in poker. This paradigm has now found applications in LLMs, where increased compute at test-time allows for more thorough exploration of possible responses, leading to improved reasoning and accuracy [11, 13, 29, 31–35]. Recent advancements, like the Deepseek-R1 model [4], further demonstrate the potential of test-time compute by leveraging reinforcement learning to refine the model's "chain-of-thought" and enhance complex reasoning abilities over longer horizons. In this work, we propose a significant scaling of the test-time compute paradigm using inductive biases derived from the scientific method to design a multi-agent framework for scientific reasoning and hypothesis generation without any additional learning techniques.

2.2 AI-driven scientific discovery

AI-driven scientific discovery represents a paradigm shift in how research is conducted across various scientific domains. Recent advancements, particularly the development of large deep learning and generative models, have cemented AI's role in scientific discovery. This is best exemplified by AlphaFold 2's remarkable progress in the grand challenge of protein structure prediction, which has revolutionized structural biology and opened new avenues for drug discovery and materials science [36]. Other notable examples include the development of novel antibiotics, protein binder design, and material discovery with AI [37–39].

Building on these successes with specialized, bespoke AI models, there has been recent work exploring the even more ambitious goal of fully integrating AI, especially modern LLM-based systems, into the complete research workflow, from initial hypothesis generation all the way to manuscript writing. This end-to-end integration represents a significant shift, presenting both unprecedented opportunities and significant challenges as the field moves beyond specialized AI tools toward realizing the potential of AI as an active collaborator, or even, as some envision, a nascent "AI scientist" [40, 41].

As an example of this shift, Liang et al. [42] directly assessed the utility of LLMs for providing feedback on research manuscripts. Through both a retrospective analysis of existing peer reviews and a prospective user study, they demonstrated the significant concordance between LLM-generated feedback and that of human reviewers. Their study, using GPT-4 [43], found that a majority of researchers perceived LLM-generated feedback as helpful, and in some instances, even more beneficial than feedback from human colleagues. However, while valuable, their work focuses solely on the feedback stage of the scientific process, leaving open the question of how LLMs might be integrated into the full research cycle, from hypothesis formation to experimental validation and manuscript writing.

Another effort embodying this shift is PaperQA2 [44], an AI agent for scientific literature search and summarization. The authors claimed to surpass PhD and postdoc researchers on multiple literature research tasks, as measured both by performance on objective benchmarks and human evaluations. While the system is a useful for synthesizing information, it does not engage in scientific reasoning for novel hypothesis generation.

HypoGeniC, a system proposed by Zhou et al. [45], tackles hypothesis generation by iteratively refining hypotheses using LLMs and a multi-armed bandit-inspired approach. The process begins with a small set of

examples, from which initial hypotheses are generated. These hypotheses are then iteratively updated through exploration and exploitation, guided by a reward function based on training accuracy. This refined set of hypotheses is subsequently used to construct an interpretable classifier. However, the method's reliance on retrospective data for evaluation means the degree to which the system can generate truly novel hypotheses remains an open question. Furthermore, the system lacks end-to-end validation beyond subjective human evaluations.

Ifargan et al. [46] present "data-to-paper", a platform that systematically guides multiple LLM and rule-based agents to generate research papers, with automated feedback mechanisms and information tracing for verification. However, the evaluations are limited to recapitulating existing peer-reviewed publications and its unclear if the system can generate truly novel, yet grounded hypothesis and research proposals.

Virtual Lab [47] is another closely related work. Here, the authors propose a team of LLM agents with a "principal investigator" LLM guiding a team of specialized LLM agents to solve a scientific problem. The LLM team receives high level human supervision. The authors demonstrate the utility of their work by leveraging Virtual Lab to design nanobody binders to recent variants of SARS-CoV-2 with experimental validation. While similar in spirit, there are significant design differences to our approach and the generality of the system remains unclear.

Boiko et al. [48] introduced "Coscientist", a multi-agent system powered by GPT-4, designed for autonomous execution of complex chemical experiments. This system integrates capabilities such as web and document searching, and code execution, to facilitate independent experimental design, planning, and execution. In addition to similar sounding names, both "Coscientist" and our system share the overarching goal of accelerating scientific discovery through AI. However, there are several important distinctions. Notably, "Coscientist" is quite narrowly focused on chemical research while ours is much broadly applicable across science. Secondly, our system has important technical innovations that lead to a self-improving system that can uncover new, original knowledge while their approach is a more vanilla-stitching of GPT-4 based agents. Finally, despite the name, "Coscientist" prioritizes a high degree of autonomy in experimental execution, directly interfacing with laboratory hardware. Our system, instead, is explicitly designed as a collaborative tool, emphasizing a "scientist-in-the-loop" approach and centers on the more cognitive aspects of the research process.

Finally, Lu et al. [40] propose "The AI Scientist", a fully automated system designed to conduct research using multiple collaborating LLM agents. These agents handle all stages of the research process, from defining research problems and conducting literature reviews to designing and executing experiments, and even writing up the results. The design shares similarities with our work—the key differences being our focus on the scaling of the test-time compute paradigm to generate high quality hypotheses and research proposals. Secondly, their proposed system has limited automated evaluations; in contrast, our work has a combination of automated, human expert and end-to-end wet lab validations. Finally, our goal is to not to automate scientific discovery, rather to build a helpful AI collaborator for scientists.

2.3 AI for biomedicine

More broadly, large AI models are increasingly demonstrating their potential in biomedical science. Both general purpose (GPT-4, Gemini) and specialized LLMs (Med-PaLM, Med-Gemini, Galactica, Tx-LLM) have shown strong performance on biomedical reasoning and question-answering benchmarks [25, 26, 49–52]. Beyond benchmarks, Med-PaLM 2, was successfully applied to identify causative murine genetic factors for traits such as diabetes, cataracts, and hearing loss [53]—an early example of hypothesis generation and LLM-assisted discovery. We have also seen the exciting development of specialized foundation and large language models trained on DNA, RNA and protein sequences with a variety of applications [54–57]. Although AI in biology and medicine often necessitates specialization, the rapid progress of frontier AI models has blurred the distinction. As these models grow in scale, data diversity, and complexity, they continue to achieve breakthroughs in areas once thought to require domain-specific AI. Our co-scientist system, with its modular multi-agent architecture, is flexibly designed to build on top of these advancements in general-purpose frontier AI models and leverage specialized AI models as tools to enhance the capabilities.

Drug repurposing is an important area of validation experiments in this work. The traditional approach to this task requires both computational and experimental approaches and a comprehensive understanding of

disease-drug interactions [17, 58]. While methods like knowledge graphs with graph convolutional networks have shown promise [59, 60], their applicability is limited by the initial knowledge graph's scope. TxGNN [61], an example of a specialized biomedical foundation model with a graph based approach, addresses "zero-shot" repurposing for novel diseases but remains dependent on the underlying knowledge graph's quality and lacks sufficient scalability and explainability. Furthermore, no end-to-end validations of the model predictions were reported in the study. In contrast, our work, leveraging state-of-the-art LLMs in the co-scientist setup, is more scalable. We report a combination of expert evaluations and wet-lab experiments to validate the system predictions.

3 Introducing the AI co-scientist

This section describes the technical details, agents, and framework comprising the co-scientist system. The co-scientist employs a multi-agent architecture built upon Gemini 2.0, integrated within an asynchronous task execution framework. This framework allows for flexible scaling of test-time compute resources, facilitating advanced scientific reasoning.

Given a research goal specified by an expert scientist in natural language, the co-scientist generates hypotheses and research proposals that adhere to the following default criteria:

- **Alignment with the provided research goal.** The generated outputs must precisely align with the research goals, preferences and constraints defined by the scientist.
- **Plausibility.** The system outputs should be free of readily apparent flaws. Any potential contradictions with prior literature or established knowledge must be explicitly stated and justified.
- **Novelty.** A key objective of the co-scientist system is to generate novel hypotheses, conjectures, and research plans grounded in prior literature, rather than simply synthesizing existing information (a capability already addressed by existing "deep research" tools [62]).
- **Testability.** The system outputs should be amenable to empirical validation within the constraints specified by the scientist.
- **Safety.** The system outputs will be controlled to prevent enabling unsafe, unethical, or harmful research.

Aside from these default criteria, the co-scientist can be configured with additional criteria, preferences, and constraints as needed. For instance, it can be configured to generate outputs in formats preferred by the researcher to improve interpretability and readability.

Throughout this section, we employ a recurring example: generating hypotheses for exploring the biological mechanisms of Amyotrophic Lateral Sclerosis (ALS) to illustrate the various components of the co-scientist system. While 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. All the examples are listed in the Appendix Section A.3.

3.1 The AI co-scientist system overview

At a high level, the co-scientist system comprises four key components:

- **Natural language 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 guide the system's progress.
- **Asynchronous task framework.** The co-scientist employs a multi-agent system where 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 utilize computational resources and iteratively improve its scientific reasoning capabilities.
- **Specialized agents.** Following inductive biases and scientific priors derived from the scientific method, the process of scientific reasoning and hypothesis generation is broken down into sub-tasks. Individual,

specialized agents, each equipped with customized instruction prompts, are designed to execute these sub-tasks. These agents operate as workers coordinated by the Supervisor agent.

- **Context memory.** In order to enable iterative computation and scientific reasoning over long time horizons, the 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 Gemini 2.0 model is the foundational LLM underpinning all agents in the co-scientist system. The specific co-scientist design was arrived at with iterative developments and is reflective of the current capabilities of the underlying LLMs.

3.2 From research goal to research plan configuration

The research goal, specified by the scientist, serves as the entry point to the co-scientist system. Leveraging the multimodal and long context capabilities of Gemini 2.0 models, the 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 (e.g., including hundreds of prior publication PDFs). The research goal may also incorporate specific constraints, attributes, and preferences related to the scientist's particular laboratory setting or field of work.

The co-scientist system 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 the 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 and 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 Appendix Figure A.9, where the goal is to develop a novel hypothesis related to phosphorylation of the Nuclear Pore Complex (NPC) as a causative mechanism for ALS [63].

Based on 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.

3.3 The specialized agents underpinning the AI co-scientist

At the core of the co-scientist system are 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. They are also equipped to interact with external tools, such as web search engines and specialized AI models, through application programming interfaces (APIs). These specialized agents are enumerated below:

- **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 novel directions, and engaging in simulated scientific debates for iterative improvement.
- **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 via literature search and review), particularly those that may be under explained.
- **Ranking agent.** An important abstraction in the co-scientist system is the notion of a tournament where different research proposals are evaluated and ranked enabling iterative improvements. The Ranking agent employs and orchestrates an Elo-based tournament [64] to assess and prioritize the

generated hypotheses at any given time. This involves pairwise comparisons, facilitated by simulated scientific debates, which allow for a nuanced evaluation of the relative merits of each proposal.

- **Proximity agent.** This agent asynchronously computes a proximity graph for generated hypotheses, enabling clustering of similar ideas, de-duplication, and efficient exploration of the hypothesis landscape.
- **Evolution agent.** The 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.
- **Meta-review agent.** This agent also enables the 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. The agent also synthesizes top-ranked hypotheses and reviews into a comprehensive research overview for review by the scientist.

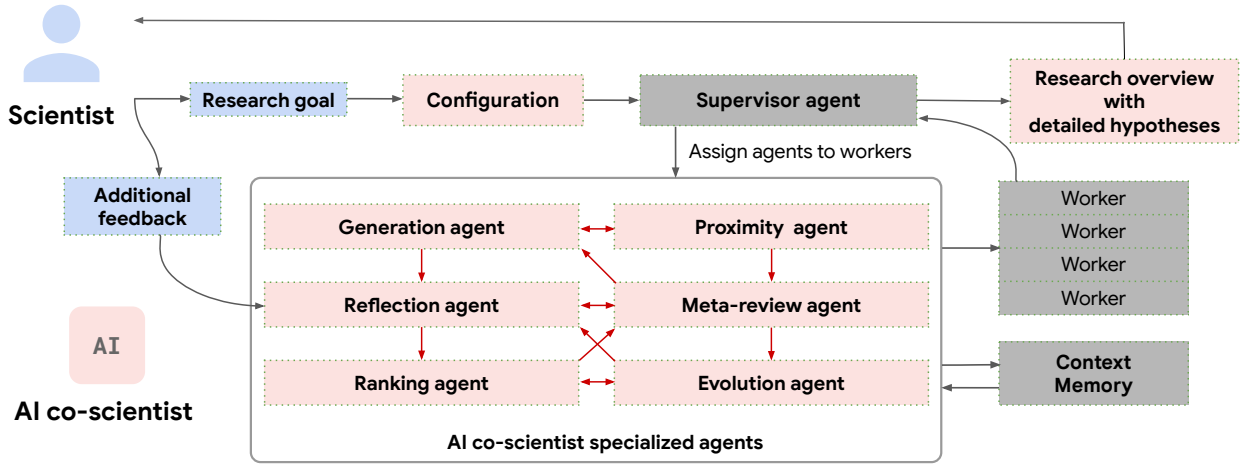

Figure 2 | The AI co-scientist multi-agent architecture design. The co-scientist accepts a natural language research goal from the user and parses this into a research plan configuration. This plan is then dispatched to the Supervisor agent which evaluates this plan to assigns weights and resources to each specialized agent and subsequently queues them as worker processes in a task queue according to these weights. The worker processes execute the queue of agent actions, and the system ultimately aggregates all information to formulate a research overview with detailed hypotheses and proposals for the scientist. The red boxes in the "The AI co-scientist specialized agents" section 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 gray arrows represent the information flow through the co-scientist system, while the red arrows represent the information feedback loop between the specialized agents.

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.

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. The Evolution, Proximity, and Meta-review agents operate on the tournament state to help 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 (e.g., generating new ideas via the Generation agent vs. improving existing ideas via the Evolution agent). Based on these statistics, the Supervisor agent then orchestrates subsequent system operations, i.e., generating new hypotheses, reviews, tournaments, and improvements to existing hypotheses, by strategically weighting and sampling the specialized agents for execution via the worker processes.

Importantly, the Meta-review agent enables feedback propagation and learning without back-propagation techniques (e.g., fine-tuning or reinforcement learning) [65]. 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 2.0 models. Through this feedback loop, the co-scientist continuously learns and improves in subsequent iterations with more compute scaling.

Finally, while our work leverages Gemini 2.0, the co-scientist framework is model-agnostic and portable to other similar models or combinations thereof. Future LLM improvements will likely enhance the co-scientist's capabilities. The multi-agent architecture of the co-scientist is depicted and summarized in Figure 2.

We now describe the mechanisms of action of the specialized agents in more detail.

3.3.1 Generation agent

The co-scientist Generation agent employs a diverse array of techniques and tools to generate novel hypotheses, such as the following:

- **Literature exploration via web search.** The agent iteratively searches the web, retrieves and reads relevant research articles, and grounds its reasoning by summarizing prior work. It then builds on this summary to generate novel hypotheses and research plans. An example prompt is given in Appendix Figure A.1.
- **Simulated scientific debates.** Here, the Generation agent simulates scientific debates among experts by employing 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 Appendix Figure A.2.
- **Iterative assumptions identification.** The agent iteratively identifies testable intermediate assumptions, which, if proven true, can lead to novel scientific discovery. These plausible assumptions and their sub-assumptions 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 Appendix Figure A.10 for the aforementioned research goal regarding explaining a basic mechanism related to ALS. The Generation agent also summarizes and categorizes each generated hypothesis, allowing scientists to quickly grasp the core ideas.

3.3.2 Reflection agent

Reviews are integral to the co-scientist's effectiveness in generating novel proposals. The Reflection agent searches relevant prior work (via 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. Effective reviews filter inaccurate and, when stipulated, non-novel hypotheses. Moreover, they also provide feedback to all other agents, driving continuous improvement. The Reflection agent employs the following types of review:

- **Initial review.** Building on the 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. For a more in-depth discussion on safety considerations see Section 6. This initial review, which doesn't use external tools like 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, leveraging 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 underlying

assumptions and reasoning. For novelty, it summarizes known aspects of the hypothesis and then judges their novelty based on existing literature. An example full novelty review is shown in Appendix Figure A.11, and an example of review critiques is in Appendix Figure A.12. A complete full review example is shown in Appendix Figure A.13.

- **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 sub-assumptions, 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 the co-scientist detect subtle errors within complex hypotheses, such as flaws in reasoning or inaccurate experimental protocols. An identified error doesn't 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 Appendix Figure A.14 for the previously introduced ALS hypothesis. We also show another example of a deep verification review via probing questions in the context of drug repurposing for AML in Appendix Figure A.15.
- **Observation review.** In addition, the Reflection agent also explores whether a given hypothesis can account for long-tail observations from prior experimental results. This review aims to determine if the hypothesis can provide insights on existing experimental findings and observed phenomena within relevant articles. For each observation, the agent assesses if 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 of the ALS hypothesis example). An example prompt to generate observations is provided in Appendix Figure A.3. An illustrative example of an observation review is provided in Appendix Figure A.16 in the context of an alternate hypothesis for explaining a mechanism of anti-microbial resistance.
- **Simulation review.** The Reflection agent also reviews hypotheses by simulating them in a step-wise fashion (e.g., simulating the mechanism of action or the proposed experiment in the proposal). This simulation allows 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 that enables them to simulate and accurately predict various scientific phenomena.
- **Recurrent/tournament review.** The Reflection agent adapts its full reviews based on the co-scientist's growing knowledge. By analyzing 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.

Additionally, the co-scientist can incorporate reviews from expert scientists to guide ranking and improvements (further discussed in Section 3.4). We aim to have the Reflection agent's comprehensive set of reviews cover the common methods scientists employ when critiquing and refining research hypotheses and proposals.

3.3.3 Ranking agent

The AI co-scientist explores numerous hypotheses and research proposals towards a research goal, necessitating a ranking mechanism to prioritize computational resources toward the most promising candidates. This task is performed by the Ranking agent. The agent uses an Elo-based tournament [64] 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 limitations [66], Elo remains a good proxy for relative ranking, and it has previously been applied to rank extracted patterns and ideas in games [67]. In the future, extensions may be considered [68]. We set the initial Elo rating of 1200 for the newly added hypothesis.

Because the tournament is computationally intensive, the Ranking agent employs several optimization strategies. Top-ranked hypotheses are compared pairwise in tournament matches through multi-turn scientific debates [69]. This mitigates ordering bias and focuses on novelty, correctness, and testability. Lower-ranked hypotheses undergo single-turn comparisons in a pairwise fashion in their tournament match. The agent concludes each comparison with a decision regarding which hypothesis is better. Appendix Figure A.4 and

Appendix Figure A.5 show example prompts. Appendix Figure A.17 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 (based on 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 favorable rankings and this informs the tournament state for subsequent iterations.

3.3.4 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 doesn't 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 allows scientists to quickly explore areas of interest and easily identify related concepts.

3.3.5 Evolution agent

The Evolution agent continuously refines and improves existing hypotheses and proposals using several approaches including:

- **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. Appendix Figure A.6 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. Appendix Figure A.7 provides an example prompt for this.

The Evolution agent generates new hypotheses; it doesn't 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 the co-scientist to iteratively combine different improvement techniques and gradually improve the quality of the results.

3.3.6 Meta-review agent

The Meta-review agent plays a crucial role in the co-scientist's feedback loop, enabling self-improvement in scientific 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 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 prevent the recurrence of common issues.

Appendix Figure A.8 provides an example prompt for the meta-review. In Appendix Figure A.18-A.19, we showcase an example of the summarized meta-review critique generated for the reviews of the previously

introduced ALS mechanism hypotheses.

Research overview generation. The Meta-review agent periodically 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 the co-scientist system and helps highlight future areas of exploration. In Appendix Figure A.20-A.21, 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 techniques [70] to adhere to common research publication and grant formats (e.g., National Institute of Health (NIH) Specific Aims Page format). We demonstrate the effectiveness of this in subsequent sections.

Research contacts identification. The Meta-review agent uses prior literature review to suggest qualified domain experts for research hypotheses 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 Appendix Figure A.22.

3.4 Expert-in-the-loop interactions with the co-scientist

The AI co-scientist empowers scientists to actively guide the system through an expert-in-the-loop design (Figure 2). Scientists can interact with the system in several ways:

- Refine the initial research goal in light of the generated hypotheses and research overview.
- Provide manual reviews of generated hypotheses (see Section 3.3.2 for other system generated review types), which the co-scientist uses to evaluate and improve the hypotheses and proposals.
- 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.
- Direct the 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, the co-scientist can prioritize generation methods that can access and synthesize it.

3.5 Tool use in AI co-scientist

The co-scientist leverages various tools during the generation, review, and improvement of hypotheses and research proposals. Web search and retrieval are primary tools, important for grounded, up-to-date hypotheses.

For research goals that explore a constrained space of possibilities (e.g., all known cell receptors of a specific type or all FDA-approved drugs), the co-scientist agents utilize domain-specific tools, such as open databases, to constrain searches and generate hypotheses. The co-scientist can also index and search a private repository of publications specified by the scientist.

Finally, the system can utilize and incorporate feedback from specialized AI models like AlphaFold. We demonstrate this qualitatively with a protein design example in the Appendix Section A.6.

4 Evaluation and Results

We now discuss the methods for evaluating the AI co-scientist system and the corresponding results. The initial evaluations aim to benchmark and verify the choice of the strategies and metrics underpinning the co-scientist. We then proceed to perform a small-scale evaluation with domain experts to assess the quality of the system.

Furthermore, to assess the practical utility of the system's novel predictions, we also perform end-to-end wet-lab validations (laboratory experiments) of the co-scientist-generated hypotheses and research proposals in three key biomedical applications: drug repurposing, discovering novel treatment targets, and elucidating the mechanisms underlying antimicrobial resistance. The varying complexity and nature of these applications enable a more comprehensive assessment of the system. Notably, all three validations involved expert-in-the-loop guidance and prioritization of experiments. These applications are summarized in Table 1.

Application	**Drug repurposing**	**Novel treatment target discovery**	**Explain mechanism of gene transfer evolution**
Challenge	Combinatorial 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

Table 1 | Three real-world applications in biomedicine for end-to-end validation of the AI co-scientist.

4.1 The Elo rating is concordant with high quality AI co-scientist results

The Elo auto-evaluation rating is a key metric that guides the self-improvement feedback loops within the co-scientist system. Therefore, it's necessary to measure and ensure higher Elo ratings correlate with higher quality results. To assess this, we analyzed the concordance between the Elo rating and the system's accuracy on the GPQA benchmark dataset. Ideally, higher Elo ratings should correlate with a higher probability of correct answers.

The GPQA dataset is a challenging, multiple-choice question answering benchmark developed by experts in biology, physics, and chemistry [71]. To ensure that the co-scientist Elo rating serves as an objective metric reflecting the validity and correctness of results from the system, we utilized questions within the GPQA diamond set, a subset of the GPQA dataset known for its high difficulty, framing each question as a research goal into our AI system to elicit responses. For each question, we first compared each co-scientist response against the ground truth answer to evaluate its correctness. Then, we categorized all generated responses across all considered questions based on their Elo rating into discrete buckets: Elo rating of 1001-1050, 1051-1100, 1101-1150, etc. in 50 point increments, until the highest rating achieved. Finally, we calculated the average accuracy for each Elo rating bucket, as the percentage of correct responses within each bucket.

We employed the underlying Gemini 2.0 models in the AI co-scientist to create a reference baseline. The reference is necessary because responses within a particular Elo rating bucket are not uniformly distributed across the GPQA questions - some of which are inherently more challenging than others. This non-uniformity could introduce bias into the analysis and potentially lead to erroneous conclusions. We therefore used the reference to generate 32 responses for each GPQA question. The fraction of correct responses from Gemini 2.0 was used as a reference accuracy on that particular question. To determine reference accuracy for a specific Elo bucket, we averaged the reference accuracy of the GPQA questions that had co-scientist responses within that bucket. We also computed the co-scientist accuracy on the GPQA diamond set by using the result with the highest Elo rating for each question and comparing it against the ground truth.

Our analysis using questions from the GPQA diamond set reveals a concordance between the Elo rating and averaged accuracy of generated co-scientist results, as depicted in Figure 3. By selecting the top-rated co-scientist result for each question, the co-scientist achieves a top-1 accuracy of 78.4%.

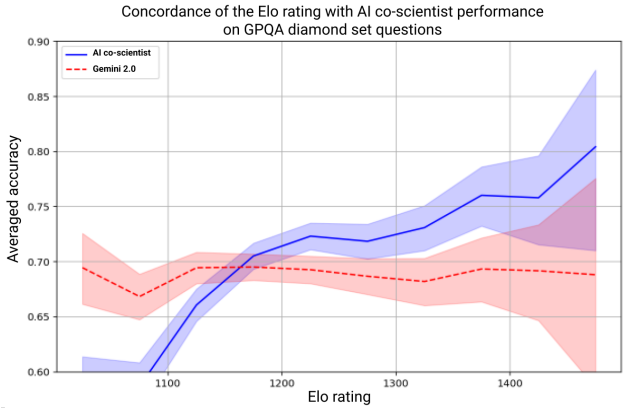

Figure 3 | Concordance of the auto-evaluation Elo metric with AI co-scientist performance on GPQA. The blue line in the figure shows the average accuracy of co-scientist responses, grouped by their Elo rating. The red line indicates the average accuracy of the corresponding reference Gemini 2.0 responses to the same set of GPQA questions, grouped by Elo rating. Note that Elo metric is auto-evaluated and not based on the ground truth.

4.2 Scaling test-time compute improves scientific reasoning of the AI co-scientist

To evaluate the effects of test-time compute scaling and the co-scientist's progress during iterative scientific reasoning and hypothesis generation, we measured the Elo ratings of the co-scientist generated hypothesis and proposals over the course of the tournament. 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 the co-scientist system until February 3, 2025.

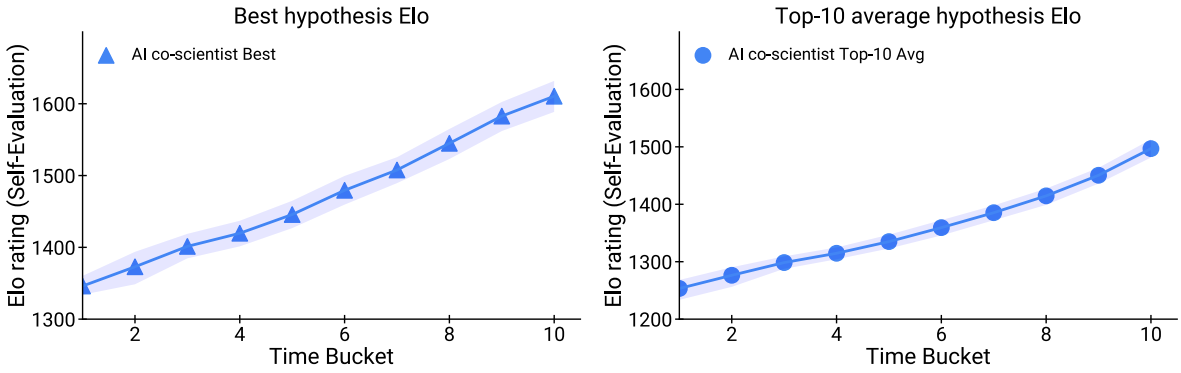

Figure 4 | Impact of scaling test-time compute on AI co-scientist as measured by Elo auto-evaluation. The co-scientist's research hypotheses and proposals were partitioned into ten temporal buckets of equal size, with the last bucket corresponding to the most recently generated results from the system. For each bucket, we determined the maximum individual Elo rating (the "best Elo") and the average Elo rating of the top 10 hypotheses across 203 unique research goals. The resulting upward performance trends, across both metrics, suggest improvements in the co-scientist result quality with scaling of test-time compute. Note that the Elo metric is auto-evaluated and not based on independent ground truth.

The 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 10 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 Figure 4, across both metrics, serves as a measure of the co-scientist's quality improvement over time. The most recent results demonstrated a significant quality enhancement compared to the initial outputs. Notably, although the Elo rating is not the direct optimization target, its progressive increase emerges from the system's self-improvement feedback loops.

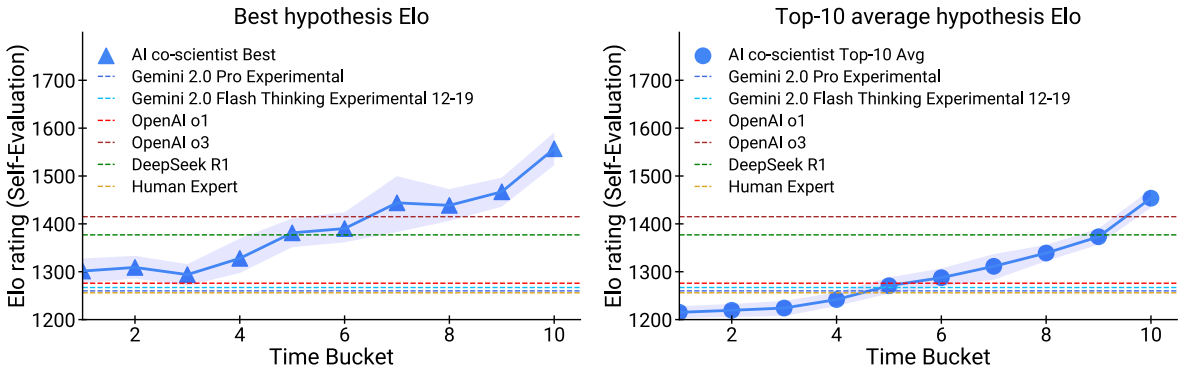

Figure 5 | Comparison of the AI co-scientist with other baselines as measured by Elo auto-evaluation on a subset of 15 challenging expert-curated research goals. The AI co-scientist's research hypotheses and proposals were partitioned into ten equal-sized temporal buckets, with the last bucket representing the most recent system-generated results. For each bucket, we calculated the maximum individual Elo rating ("best Elo") and the average Elo rating of the top 10 hypotheses across 15 challenging research goals, curated by biomedical experts. These metrics were then compared to the performance of several baselines: Gemini 2.0 Pro Experimental, Gemini 2.0 Flash Thinking Experimental 12-19, OpenAI o1, OpenAI o3-mini-high, DeepSeek R1, and the "best guess" hypotheses provided by the expert curators. These baseline performances are represented by colored dashed lines. On this expert-curated subset of research goals, we consistently observed upward performance trends with increased test-time compute. Furthermore, the co-scientist significantly outperformed both the human experts and the other state-of-the-art LLM baselines, as measured by the Elo metric. It's important to note that the Elo metric is auto-evaluated and not based on independent ground truth. This auto-evaluation metric may favour results with attributes that do not necessarily align with scientists' preferences, quality, or accuracy.

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. These experts held doctoral degrees in biological or life science disciplines and were actively working as postdoctoral researchers or faculty members at academic biomedical research institutes. 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—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 "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 Figure 4 were consistent within this subset. Furthermore, as shown in Figure 5, the co-scientist surpassed the other frontier LLMs and reasoning models in Elo rating with increased computational resources for iterative improvement. Notably, newer reasoning models, such as OpenAI o3-mini-high and DeepSeek R1, demonstrated competitive performance while requiring significantly 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 result quality of the co-scientist system. Its worth noting again that the co-scientist architecture is model agnostic and is likely to benefit from further advancements in frontier and reasoning LLMs.

Building upon the co-scientist system's ability to combine, refine and improve research hypotheses and proposals iteratively, we investigated its potential to improve upon expert "best guess" solutions. Consistent with our previous observations, the co-scientist demonstrated the capacity to enhance expert's "best guess" solutions over time, as evidenced by the Elo metric in Figure 6. Notably, the improvement trends initially mirrored those of the co-scientist's autonomously generated solutions but subsequently surpassed them. While this is a preliminary finding requiring further validation, it suggests a promising avenue for capable AI systems, such as the co-scientist, to augment and accelerate the work of expert scientists.

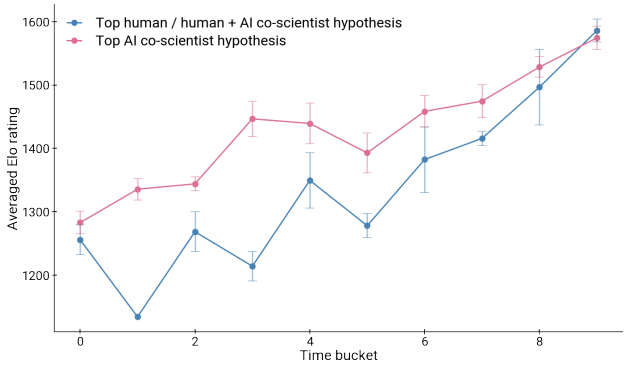

Figure 6 | AI-augmented expertise with the co-scientist through Elo-based auto-evaluation. Through its self-improvement process, the co-scientist refines and enhances expert "best guess" solutions over time, as measured by the Elo rating on a subset of 15 curated research goals. It is important to note that the Elo metric is auto-evaluated and not based on independent ground truth.

4.3 Experts consider the AI co-scientist results to be potentially novel and impactful

To obtain expert feedback and assess preferences, we conducted a small-scale expert evaluation on 11 of the 15 previously curated research goals. We asked the experts who curated the research goals to assess outputs from the AI 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 5-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 significant 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, outputs generated by the AI co-scientist were most preferred and rated higher in novelty and impact axes compared to the other baseline models. Specifically, the 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 Figure 7. 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 Figure 5 and Figure 7.

We also conducted the same preference ranking evaluation between co-scientist and other LLM and reasoning model baselines using the 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. The co-scientist outputs were the most preferred by both the o3-mini, o1 and Gemini 2.0 Pro Experimental models as shown in (Figure 8). Due to the small scale of these evaluations, further large-scale studies are necessary for any reliable conclusions. We present a more comprehensive clinical expert evaluation focused on co-scientist proposals for drug repurposing formatted in the NIH Specific Aims Page format in Section 4.5.1.

4.4 Safety evaluation of the AI co-scientist using adversarial research goals

The AI co-scientist is designed to empower scientists and accelerate research. However, it's crucial to ensure the system is designed with robust safety principles, given the potential for misuse. This includes addressing dangerous research goals, dual-use objectives, scenarios where safe goals lead to unsafe hypotheses, misleading claims, and inherent biases. While this topic requires extensive investigation beyond the scope of this work, we employed adversarial testing strategies to conduct a preliminary safety analysis of the system. Specifically,

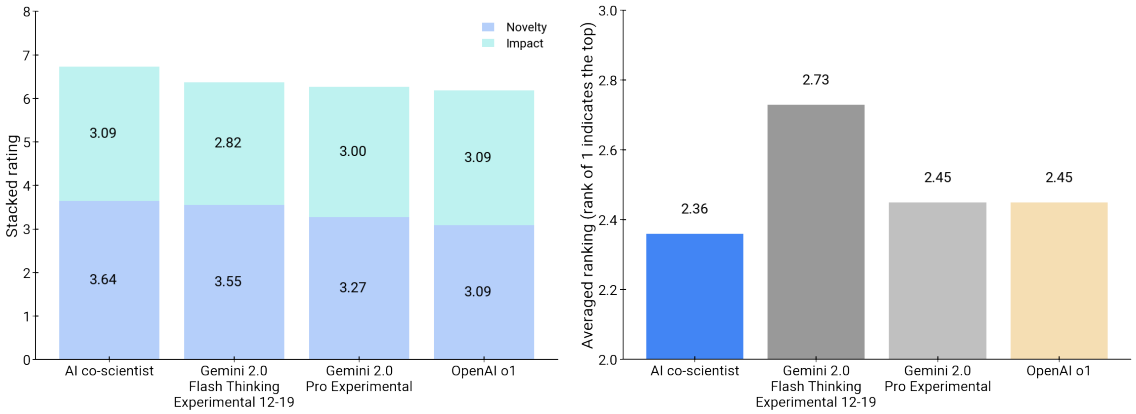

Figure 7 | Expert evaluation of AI co-scientist and other LLM baselines. Left: Average expert ratings on novelty and impact of the model responses across 11 expert curated research goals. Higher numbers indicate better ratings (1-5). Right: Average expert preference ranking of the results across 11 expert curated research goals generated by AI co-scientist, Gemini 2.0 Flash Thinking Experimental 12-19, Gemini 2.0 Pro Experimental, and OpenAI o1, respectively. Lower numbers indicate better rankings (1-4). The human expert preferences also appear to be concordant with relative Elo ratings as can be inferred from Figure 5. At the same time, its worth noting that these preferences and ratings reflect subjective expert assessments, not objective ground truth.

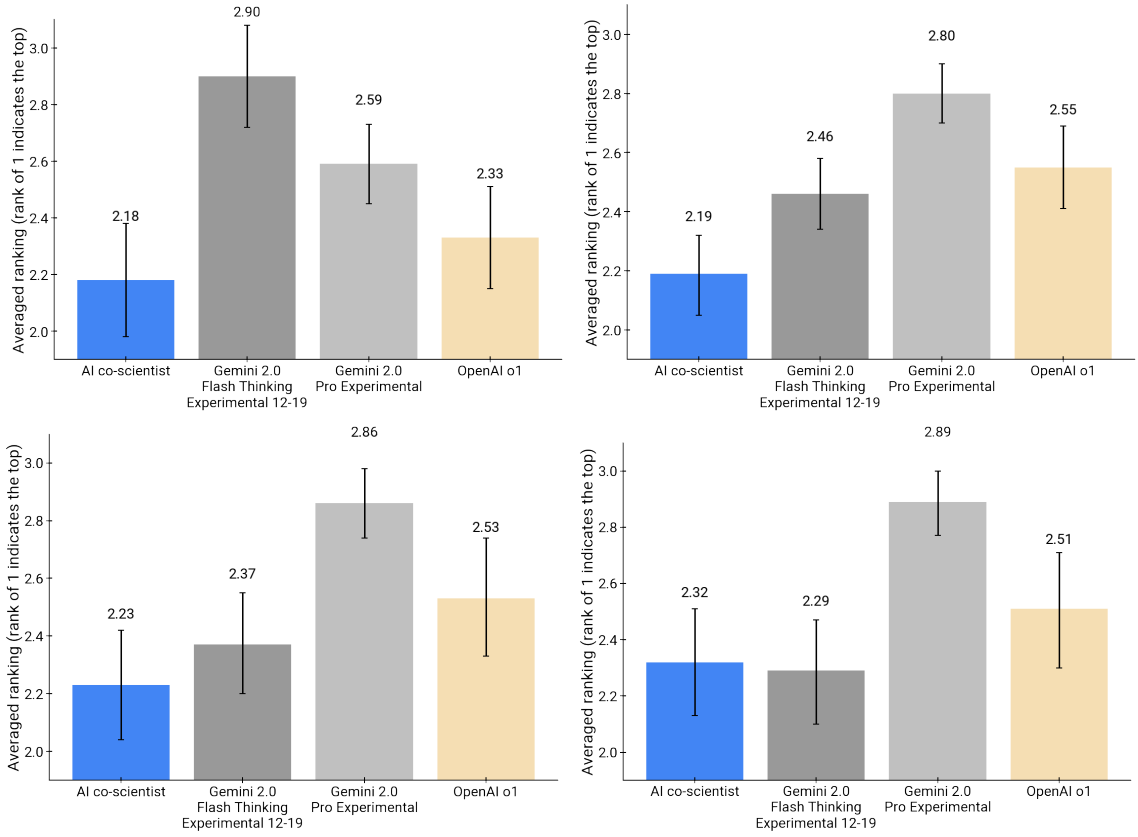

Figure 8 | LLM preference ranking auto-evaluation of AI co-scientist and other baselines. Averaged preference ranking of results across 11 expert curated research goals generated by AI co-scientist, Gemini 2.0 Flash Thinking Experimental 12-19, Gemini 2.0 Pro Experimental, and OpenAI o1, using four different 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.

we curated a set of 1200 adversarial examples, ranging in complexity, across 40 biomedical and scientific topics using frontier LLMs. We then evaluated whether the AI co-scientist could robustly reject these research goals. In this preliminary analysis, the system successfully passed all checks. Given the sensitive nature of these adversarial research goals, we will not be publicly releasing the dataset, but it can be made available upon request. Collectively, the benchmark, automated, and expert evaluations presented in this section provide compelling evidence of the system's strong capabilities.

4.5 Drug repurposing with the AI co-scientist

As previously noted, a rigorous assessment of a system's ability to generate novel hypotheses and predictions for complex research problems necessitates end-to-end validation through wet-lab experiments. However, due to the challenging, time-consuming, and resource-intensive nature of such endeavors, large-scale experimental validation is infeasible. Instead, we strategically selected diverse yet critical biomedical topics to serve as a strong benchmark for the end-to-end system evaluation. Detailed descriptions of these topics follow. Importantly, all three experimental validations were conducted in collaboration with expert scientists, who provided guidance to the co-scientist and prioritized wet-lab experiments.

We begin the discussion of the end-to-end validation of the AI co-scientist with a drug repurposing application. As introduced earlier, drug repurposing is the process of identifying novel therapeutic indications for existing, approved drugs beyond their original use. This approach can accelerate the discovery 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 combinatorial search problem involving a large but finite set of drug-disease pairs as noted in Table 1.

Given the 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 capabilities. The system is general-purpose, capable of providing highly detailed and explainable predictions across all known drug-disease pairs. Here, we focused on the computational biology and wet-lab validation of our co-scientist system in the area of drug repurposing for cancer treatment.

We initially investigated drug-cancer pairs with existing preclinical evidence to validate the plausibility of the hypotheses and predictions generated by the co-scientist (Section 4.5.1), before expanding to completely novel drug repurposing hypotheses (Section 4.5.2). The validation of the co-scientist's predictions was performed using a multi-faceted approach, incorporating computational biology analyses, oncologist expert feedback, and *in vitro* wet-lab experiments using cancer cell lines.

4.5.1 The AI co-scientist suggests plausible drug repurposing candidates as rated by experts

We constrained the AI co-scientist to explore potential repurposing hypotheses from a curated list of 2300 approved drugs across 33 cancer types (Appendix Section A.4.1). To achieve this, we modified the prompts used in the Generation and Ranking agent stages to ensure hypotheses generation in this constrained search space; however, the core co-scientist logic remained unchanged. When formulating the research goal for the co-scientist, we explicitly emphasized the following preferences related to drug repurposing:

- Elucidate the known mechanisms of action and impacted biological pathways of the drug.
- Identify potential diseases or cancer types that could be treatment targets for the drug.
- Explain the potential mechanisms by which the drug could exert therapeutic effects.
- Propose alternative mechanisms of action through which the drug might function in the proposed therapeutic context.
- Identify the diseases / cancers for which the drug is currently approved.
- List the most promising disease / cancer type candidates for repurposing.
- Discuss prior research and challenges associated with repurposing the drug.

For each drug-cancer pair, we also extracted the Cancer Dependency Map (DepMap) probability of dependency ("DepMap score") [72] (Appendix Section A.4.2). The DepMap score represents the probability of essentiality for a gene in a given cancer cell lines. We ranked all drug-cancer pairs using a combined metric of the co-scientist review score (ranging from 1 to 5) and the DepMap score (ranging from 0.0 to 1.0). To prioritize

the most relevant hypotheses for expert review, we selected only pairs where the co-scientist review score ≥ 4 and the DepMap score ≥ 0.99. Note that the DepMap score is primarily meant to function as a sanity check and filter out obviously incorrect candidates but is unlikely to be predictive of efficacy.

Expert oncologists then reviewed the top-ranked drug-cancer pairs, provided feedback, and selected promising repurposing candidates for *in vitro* wet-lab validation (Appendix Section A.4.3).

Clinical expert evaluation of drug repurposing proposals in NIH Specific Aims Page format. To rigorously evaluate whether the co-scientist-generated hypotheses for drug repurposing fulfill the needs of physicians and scientists, we restructured the co-scientist hypotheses into the NIH-style grant proposal Specific Aims Page (examples in Appendix Figure A.26-A.31), and asked a team of six expert hematologists & oncologists to evaluate the specific aims.

The NIH Specific Aims Page format follows a standard structure, including disease description, unmet need, proposed solutions, and specific aims. This format was selected because it provides a standardized framework that is widely recognized in the research community, allowing for systematic presentation of complex scientific topics in a manner that facilitates rigorous peer review and enables efficient assessment of scientific merit. The specific aims, which outline the overarching goal, hypothesis, and rationale, requires extensive scientific expertise, comprehensive literature analysis, and robust domain knowledge. We generated cancer drug repurposing hypotheses derived from the co-scientist in the format of NIH Specific Aims Page with additional constrained decoding and self-critique stages to ensure format consistency. The resulting format contextualizes proposed repurposing candidates within known mechanisms based on current literature and then extrapolates to a new disease state. An expert oncologist methodically evaluated and excluded hypotheses that were deemed clinically implausible or had limited potential for successful translation, as well as those falling outside the expertise of the assembled specialist evaluators. This initial screening process employed multiple evidence-based criteria including: (1) pharmacological mechanism incompatibility with tumor biology; (2) unfavorable pharmacokinetic profiles for oncological applications; (3) prohibitive toxicity profiles documented in prior clinical use; (4) confounding effects where apparent survival benefits were attributable to improved management of treatment-related morbidity rather than direct anti-neoplastic activity; and (5) insufficient preclinical evidence supporting antitumor efficacy at clinically achievable concentrations. For example, bisphosphonate agents like pamidronate, while associated with improved outcomes in observational studies of patients with bone metastases, were excluded after critical evaluation revealed their benefits stemmed primarily from reduction of skeletal-related events (such as pathological fractures, spinal cord compression, and bone pain requiring radiation) rather than from disease modifying activity of the drug-candidate.

Six board-certified hematologists & oncologists from a single institution - including four domain-specific oncologists specializing in gastrointestinal (GI), breast, gynecologic (GYN) and genitourinary (GU) cancers and two general hematologist & oncologists, with an average of eight years of clinical experience - evaluated 78 unique drug repurposing hypotheses presented in the NIH Specific Aims Page format (for specific indication distribution and counts, see Appendix Section A.5.1).

The expert raters evaluated the generated Specific Aims based on a modified NIH grant proposal evaluation rubric, consisting of 15 axes focusing on (1) importance of research (significance and innovation) and (2) approach (rigor and feasibility). The raters indicated their agreement level using a five-point scale: "Strongly Agree", "Agree", "Neutral", "Disagree", and "Strongly Disagree". For each axis, we included several questions covering different aspects of the NIH evaluation criteria. The evaluation rubric is further detailed in the Appendix Section A.5.2. Specifically, we ask raters to focus on evaluating the clinical relevance and potential for clinical translation, and not for translational capacity or the design of clinical trials.

We observed that expert raters consistently assigned high ratings ("Strongly Agree" or "Agree") to the Specific Aims proposed by the co-scientist across various evaluation criteria (Figure 9). Of note, the favorable assessments of co-scientist-generated hypotheses may be partially attributed to expert pre-screening, wherein a clinician eliminated non-viable candidates prior to expert evaluation. Three examples of generated Specific Aims and their respective expert review ratings are detailed in Appendix Figure A.26-A.31.

The generated Specific Aims were assessed by clinical hematologists & oncologists from a single-center, which might bias the interpretation of the evaluation results, as it may introduce institutional perspectives shaped by local practice patterns, clinical experiences, and research frameworks unique to that setting. While some

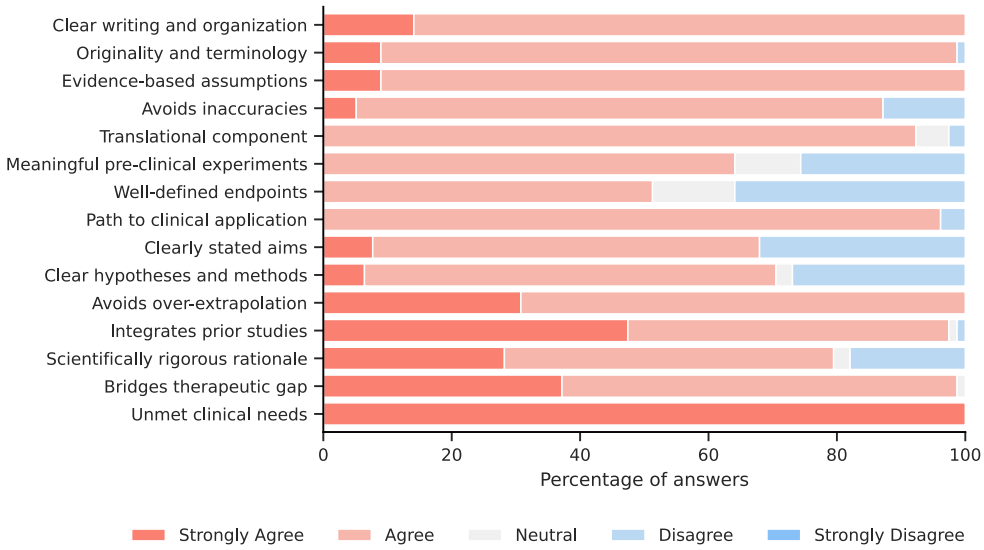

Figure 9 | Clinical expert evaluation for the co-scientist generated drug repurposing hypotheses in the NIH Specific Aims Page format. Six expert hematologists & oncologists reviewed 78 drug repurposing research proposals, which the co-scientist had formatted as NIH Specific Aims Pages. The evaluation followed an adapted NIH grant proposal evaluation rubric, detailed in Appendix Section A.5.2. Overall, the oncologists judged the Specific Aims proposals from the AI co-scientist to be of high quality across all axes of the rubric.

Specific Aims may be supported by preclinical data, it is important to note that none of the proposed drug candidates have undergone randomized phase III clinical trials necessary to establish efficacy and secure regulatory approval for repurposing to a new indication.

4.5.2 The AI co-scientist identifies novel drug repurposing candidates for acute myeloid leukemia

Building upon the positive feedback from clinical experts, we conducted *in vitro* wet-lab validation experiments for drug repurposing hypotheses generated by the co-scientist for acute myeloid leukemia (AML). AML is an aggressive and relatively rare blood cancer characterized by the rapid proliferation of abnormal white blood cells (myeloblasts) in the bone marrow, which displaces healthy blood cells. We focused on this indication due to its aggressive nature and the limited availability of effective therapeutic interventions [73].

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. The simplicity in experimental design is purposeful; our focus is on evaluating the merit of AI co-scientist generated hypotheses rather than developing novel laboratory techniques. Positive results from these experiments should be interpreted as preliminary evidence warranting further investigation through comprehensive pre-clinical studies (e.g., *in vivo* models) and potentially clinical evaluation.

It is important to emphasize that these wet-lab experiments function as a viability checkpoint in the drug repurposing pipeline - not as a replacement for the rigorous pre-clinical and clinical assessment typically required for therapeutic validation. They provide an efficient biological reality check that helps bridge the gap between computational predictions and potential clinical applications, allowing us to rapidly triage AI-generated hypotheses before committing to more resource-intensive validation studies necessary for clinical translation.

Drug repurposing candidate selection process for acute myeloid leukemia. The candidate selection for wet-lab experiments was performed with meticulous expert oversight. Thirty top-ranked drug candidates hypotheses were shared with expert oncologists (an example detailed co-scientist output is provided in Section A.5.4). The experts evaluated the hypotheses, selecting drug candidates based on their potential to modulate key molecular signaling pathways associated with disease progression and resistance. Note that

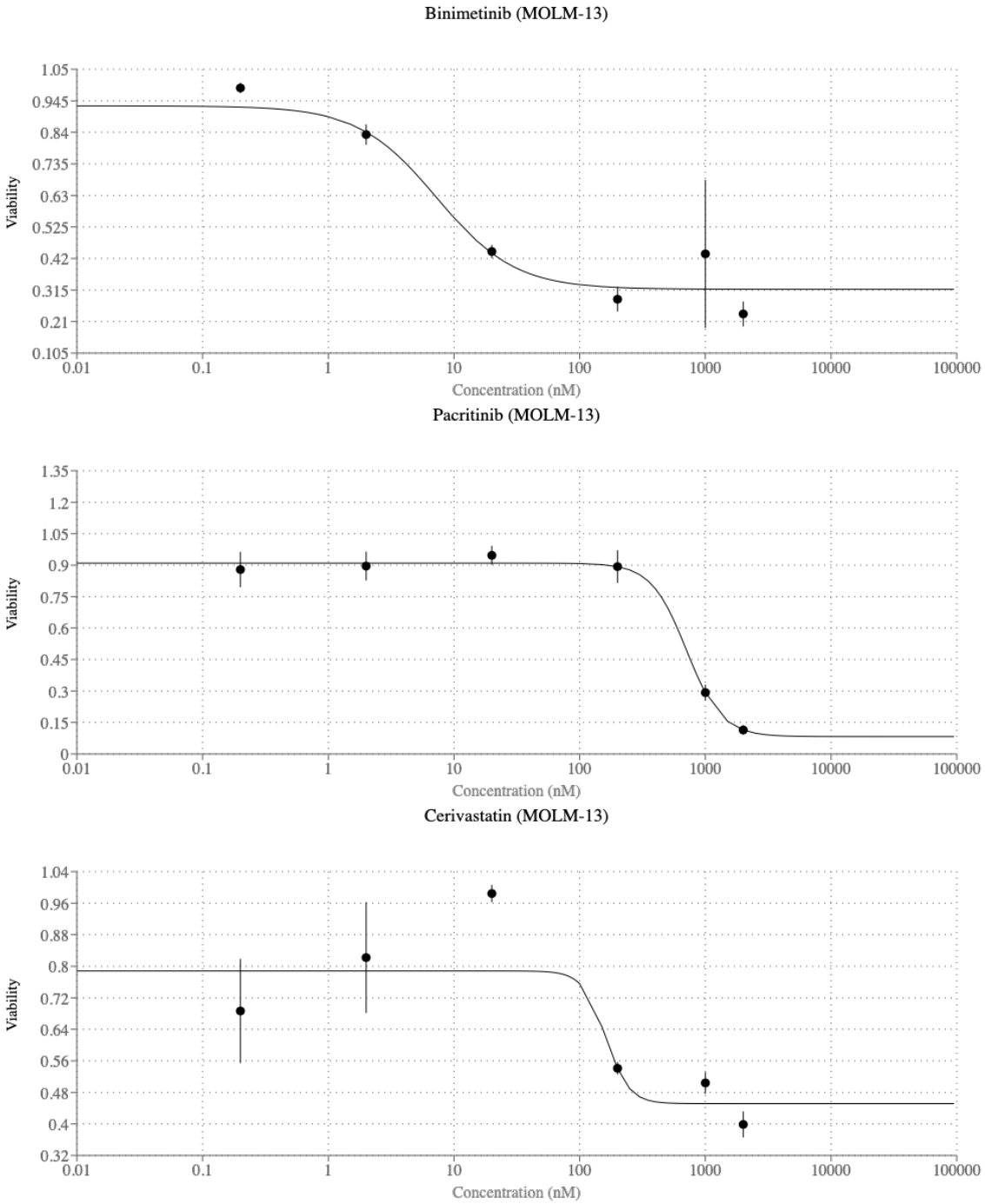

Figure 10 | Dose response curve of the expert selected repurposing drugs with existing evidence. Binimetinib, Pacritinib, and Cerivastatin inhibit MOLM-13 cell viability. X-axis is the drug concentration (nM), and Y-axis is normalized cell viability (arbitrary unit). Lower cell viability indicates that the selected drug has a stronger inhibition on AML cells.

we did not preclude any hypotheses based on whether single-agent therapy has been studied before, or if the phase III has been approved. The primary selection criterion favored compounds with multi-pathway activity, specifically those influencing dysregulated inflammatory signaling, metabolic reprogramming, and aberrant cell proliferation. Emerging research indicates that these shared biological processes play a critical role in relapse and treatment resistance [74]. Candidates were also chosen based on preclinical mechanistic insights and their relevance to AML biology, including their hypothesized effects on leukemic cell survival,

microenvironment interactions, and resistance mechanisms.

Based on potential mechanisms of action, five drug repurposing candidates—Binimetinib, Pacritinib, Cerivastatin, Pravastatin, and Dimethyl fumarate (DMF)—were selected for further wet-lab validation in AML.

Briefly, Binimetinib is an inhibitor of MEK1 and MEK2, key kinases in the RAS–RAF–MEK–ERK signaling pathway. By inhibiting MEK1/2, binimetinib prevents the activation of ERK (extracellular signal-regulated kinase), thereby blocking downstream signaling that promotes cell proliferation and survival [75]. Although RAS mutations typically emerge as late events in AML pathogenesis, Binimetinib was included to investigate its potential to modulate RAS-MEK-ERK signaling in treatment-naive AML, where baseline expression levels of this pathway can influence sensitivity to conventional chemotherapeutic agents [76].

Pacritinib is an oral tyrosine kinase inhibitor that selectively targets JAK2 and FLT3 kinases [77]. By blocking JAK2's kinase activity, pacritinib suppresses the overactive JAK-STAT signaling that drives pathogenic cell proliferation and cytokine production in diseases such as myelofibrosis. It was selected for repurposing due to the dual inhibition of growth signaling pathways: the JAK2/STAT pathway, critical in hematopoietic cell growth and inflammatory signaling, and FLT3-driven proliferative signaling that regulates leukemic cell survival and also facilitates the development of escape pathways to targeted therapies [78].

Dimethyl fumarate (DMF) is an immunomodulatory drug that activates the Nrf2 (nuclear factor erythroid 2–related factor 2) pathway via covalent modification of the cysteine residues on Keap1, the cytosolic protein that normally binds Nrf2 and targets it for degradation. By oxidizing or alkylating the thiol groups of Keap1, DMF destabilizes the Keap1–Nrf2 complex, allowing Nrf2 to escape ubiquitination and translocate into the nucleus. In parallel, DMF also inhibits NFκB mediated transcription and was chosen for repurposing due to clinically relevant activity of NFκB in AML [79, 80].

Finally, the statins (Cerivastatin and Pravastatin) were selected for their potential to induce metabolic reprogramming and directly modulate vesicular transport mechanisms in rapidly proliferating cells [81].

Laboratory *in vitro* validation of expert-selected drugs with existing evidence. Of the five drugs tested, Binimetinib, Pacritinib, and Cerivastatin demonstrated inhibition of cell viability (Figure 10). Notably, Binimetinib, which is already approved for the treatment of metastatic melanoma, exhibited an IC50 as low as 7 nM in AML cell lines (Figure 10 and Appendix Figure A.24). This result shows that the drugs proposed by the co-scientist hold promise as clinically viable drug repurposing candidates. Moreover, this opens the possibility that the co-scientist may be able to expand its hypotheses to novel drug repurposing candidates.

The AI co-scientist proposal of novel drug repurposing candidates for acute myeloid leukemia. We aimed to demonstrate the co-scientist's capacity to autonomously propose novel drug repurposing candidates without oversight. Towards this, the system was directed to generate a ranked list of repurposing candidates for AML, including drugs that were not previously repurposed for the target indication and without any prior preclinical evidence. Specifically, we tasked the co-scientist with generating potential novel drug repurposing hypotheses for AML without explicitly relying on additional external inputs, such as the DepMap scores or human expert feedback. We then determined if these novel candidates suggested by co-scientist could be validated in the laboratory, and may therefore have the potential to be repurposed for AML.

For *in vitro* laboratory validation of novel repurposing drugs, the domain experts selected the top 3 candidates from the ranked list, using the criteria that no prior preclinical or clinical data existed with respect to their use to treat AML - Nanvuranlat, KIRA6, and Leflunomide.

To demonstrate the hypothesis and the rationale given by the AI co-scientist for these drug repurposing candidates, the detailed AI co-scientist output, including the novelty review, is provided in Section A.5.4 for KIRA6. As can be the seen, the system identifies that targeting IRE1α in the context of AML has been explored before but not with the specific drug, KIRA6. 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 three different AML cell lines, KG-1, MOLM-13, and HL-60 cells (Figure 11). IC50s of KIRA6 were all in nM range, but significantly more effective in KG-1 cells, which had an IC50 of 13 nM, compared to MOLM-13 and HL60 cells, which had IC50s of 517 nM and 817 nM, respectively. Thus, co-scientist was able to suggest

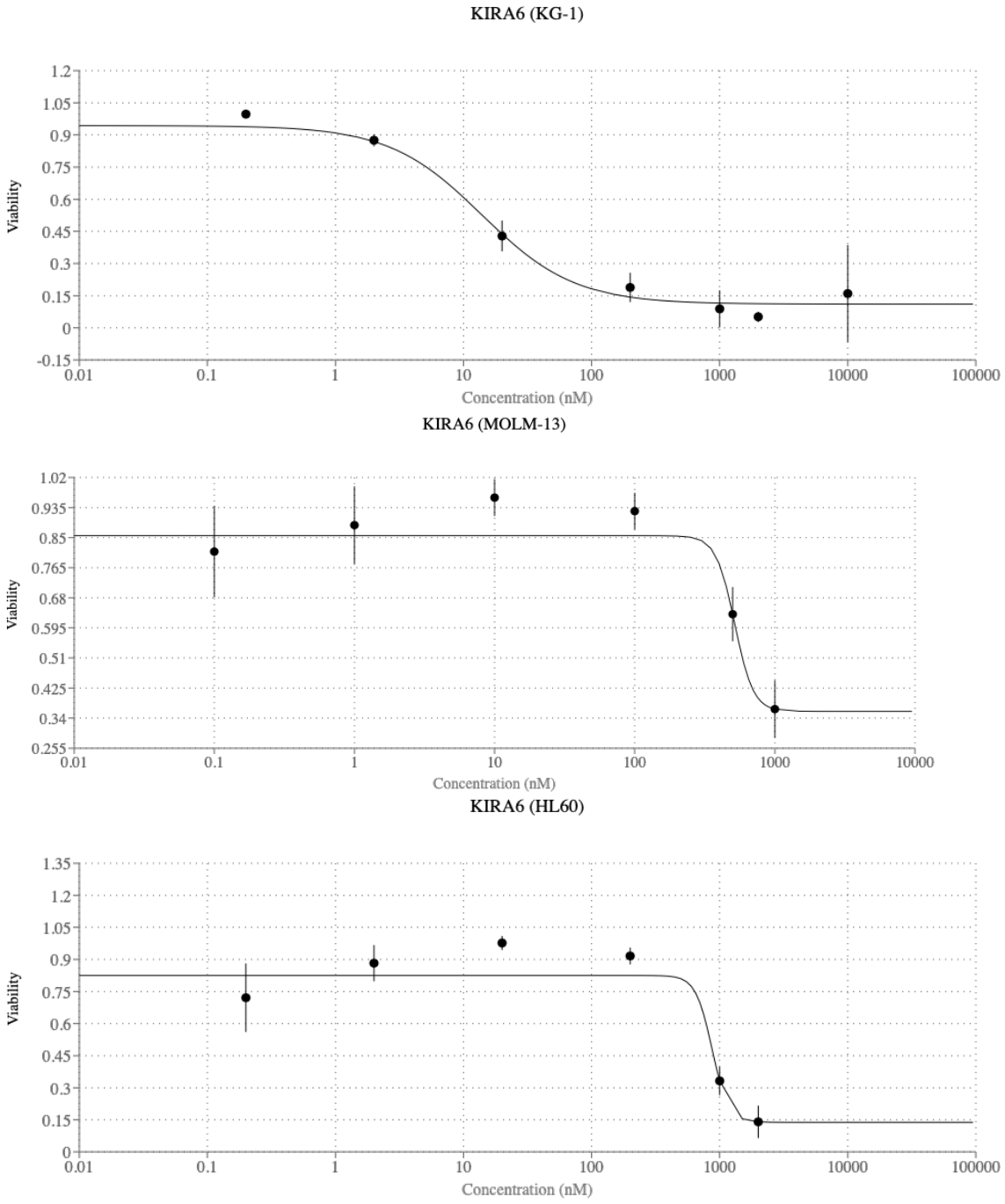

Figure 11 | Dose response curve of novel drug repurposing candidate for AML. KIRA6 activity inhibiting KG-1, MOLM-13, and HL-60 cell viability all in nM range of drug concentration. X-axis is the drug concentration (nM), and Y-axis is normalized cell viability (arbitrary unit). Lower cell viability indicates that the selected drug has a stronger inhibition on AML cells.

a novel candidate for drug repurposing for AML, beyond those that may have been selected through other existing approaches and expert sources. This suggests that the co-scientist system may therefore be capable of generating new, promising hypotheses for researchers to investigate, that may in the future bring new treatments to patients for complex and challenging diseases such as AML.

Translating these insights from co-scientist's drug repurposing hypotheses 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 hematological malignancies, the tumor microenvironment and systemic interactions may introduce unforeseen resistance mechanisms, further complicating translation from hypothesis to clinical benefit.

4.6 The AI co-scientist uncovers novel therapeutic targets for liver fibrosis

Liver fibrosis is a severe disease that can progress to liver failure and hepatocellular carcinoma, which has few treatment options due to the limitations of available animal and *in vitro* models. However, a recently developed method for producing human hepatic organoids coupled with a live cell imaging system for liver fibrosis provides a new avenue for identification of new treatments for liver fibrosis [82–84]. The AI co-scientist was asked to produce experimentally testable hypotheses concerning the role of epigenetic alterations in liver fibrosis ("A Novel Hypothesis Regarding Myofibroblast Generation in Liver Fibrosis"); and to identify drugs targeting epigenetic modifiers that could be used for treatment of liver fibrosis.

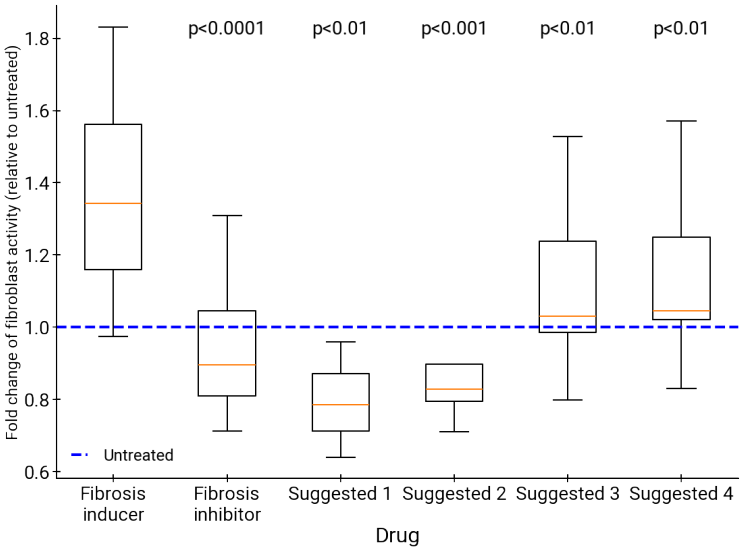

Figure 12 | The co-scientist discovers the novel treatment targets for liver fibrosis. Four drugs (Suggested 1-4) based on three epigenetic targets suggested by the co-scientist decrease the fold change of the fibroblast activity. The experiments were conducted in the human hepatic organoid system, the fibroblast activity was measured by the percentage fold change of fluorescence labelled on the myofibroblast. 'Untreated' indicates the normal control, the fibrosis inducer is the fibrogenic agent TGF-β, the fibrosis inhibitor indicates the fibrogenic stimulated myofibroblast treated with the fibrogenic agent inhibitor (i.e., TGF-β inhibitor), and four suggested drugs indicate the fibrogenic stimulated myofibroblast treated with each of the four drugs based on three AI co-scientist suggested epigenetic targets. The red line in each box is the median fold change of the group. The p-value indicates the statistical significance between the fibrosis inducer group and the given group. These results will be further detailed in an upcoming report.

The experts selected three (from fifteen) top-ranked co-scientist generated research hypotheses with a comprehensive research proposal (i.e., experimental design, evaluation methodology, and anticipated results) for exploring the role of epigenetic modifications in liver fibrosis. The co-scientist identified three novel epigenetic modifiers with supporting preclinical evidence that could be targeted by existing agents and provide new treatments for liver fibrosis. Drugs targeting two of the three epigenetic modifiers exhibited significant anti-fibrotic activity in hepatic organoids without causing cellular toxicity (Figure 12). Since one of them is an FDA-approved drug for another indication, this creates an opportunity to re-purpose a drug for treatment of liver fibrosis. These results will be detailed in an upcoming technical report.

4.7 The AI co-scientist recapitulates a breakthrough in antimicrobial resistance

Understanding the mechanisms of antibiotic resistance is crucial for researchers to develop effective treatments against infectious diseases. We focused on capsid-forming phage-inducible chromosomal islands (cf-PICIs), which play a pivotal role in antibiotic resistance. These mobile genetic elements, unlike typical phages and other PICIs, possess a remarkable ability to transfer between diverse bacterial species, carrying with them virulence and antibiotic resistance genes. We sought to understand the evolutionary rationale behind the existence of cf-PICIs across multiple bacterial species in order to develop solutions to combat antimicrobial resistance.

The primary objective was to leverage the AI co-scientist to generate a research proposal aimed at elucidating the molecular mechanisms of bacterial evolution underlying the broad host range of cf-PICIs and developing strategies to curb the spread of antibiotic resistance. We specifically focused on the observation that identical cf-PICIs, such as PICIEc1 and PICIKp1, are found in clinically relevant bacterial species, including WHO priority pathogens like *Escherichia coli* and *Klebsiella pneumoniae*.

In a co-timed study [85] currently undergoing the peer-review process at an established journal in the field, genomic and experimental studies revealed a novel mechanism explaining how identical cf-PICIs can be found in different bacterial species. Knowing the answer to that question (but without it yet being available in the public domain), we investigated whether the co-scientist could independently discover the same, or similar, research hypotheses. We provided the co-scientist with a single-page document containing general information, including a brief background on phage satellites and two relevant research articles. The first paper described the original discovery of cf-PICIs [86], and the second paper introduced a computational technique for identifying phage satellites in bacterial genomes [87]. We then challenged co-scientist to address the question of why cf-PICIs, but not other types of PICIs or satellites, are readily found across diverse bacterial species, and what mechanism underlies this phenomenon.

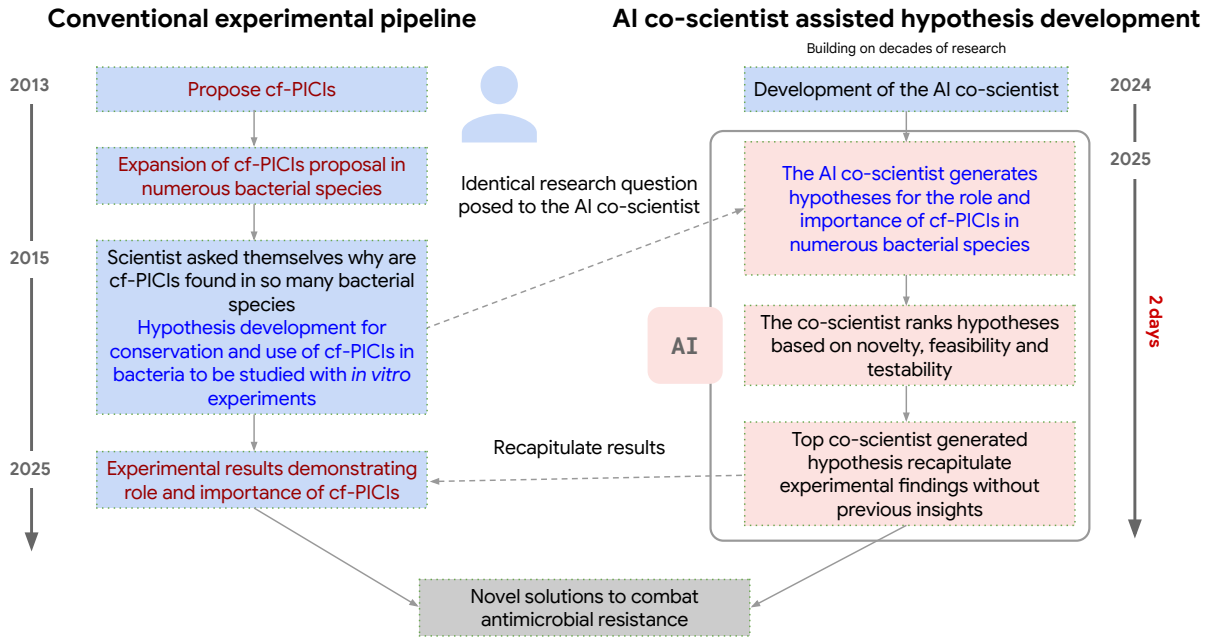

Figure 13 | Timeline of the conventional experimental validation and the AI co-scientist-assisted hypothesis development for capsid-forming phage-inducible chromosomal islands (cf-PICIs), key to antibiotic resistance (AMR). Blue box: Scientist inputs. Red box: The AI co-scientist system. Red text: wet lab experimental setting. Blue text: research hypothesis generation.

The co-scientist independently and accurately proposed a groundbreaking hypothesis—that cf-PICIs elements interact with diverse phage tails to expand their host range—as its top-ranked suggestion [21]. This finding was experimentally validated in the independent research study, which was unknown to the co-scientist during

hypothesis generation [20]. Its worth noting that while the co-scientist generated this hypothesis in just two days, it was building on decades of research and had access to all prior open access literature on this topic.

(Figure 13). Specifically, the co-scientist suggested the following research topics for the given research goal regarding cf-PICIs:

- **Capsid-tail interactions.** Investigate the interactions between cf-PICI capsids and a broad range of helper phage tails. This topic aligns precisely with the unpublished manuscript's primary finding: that cf-PICIs interact with tails from different phages to expand their host range, a process mediated by cf-PICI-encoded adaptor and connector proteins.
- **Integration mechanisms.** Examine the mechanisms by which cf-PICIs integrate into the genomes of diverse bacterial species.
- **Entry mechanisms.** Explore alternative cf-PICI entry mechanisms beyond traditional phage receptor recognition.
- **Helper phage and environmental factors.** Investigate the role of helper phages and broader ecological factors in cf-PICI transfer.
- **Alternative transfer and stabilization mechanisms.** Explore other potential transfer mechanisms, such as conjugation, extracellular vesicles, and unique stabilization strategies, that might contribute to cf-PICI's broad host range.

The convergence of the conventional and AI co-scientist approaches on the same novel finding underscores the potential of the co-scientist to augment, complement and accelerate scientific endeavors (Figure 13). Further results and comprehensive details are available in the companion report [21].

5 Limitations

We are encouraged by the early promise of the AI co-scientist evaluations, which highlight its potential to augment scientific research. However, the system has several limitations. Responsible innovation necessitates thoughtful consideration of these alongside the potential impacts to researchers and scientific research.

Limitations with literature search, reviews and reasoning. The reviews undertaken by the AI co-scientist system may miss critical prior works due to reliance on open-access literature. In the presented work, the AI co-scientist does not access the entire published literature due to compliance with license or access restrictions where applicable. The system may also omit consideration of prior work on occasions where it has incorrectly reasoned that the work is not relevant.

Lack of access to negative results data. The AI co-scientist system's use of only open published literature means it likely has limited access to negative experimental results or records of failed experiments. It is known that such data may be more rarely published than positive results, yet experienced scientists working in the field may nonetheless possess and utilize this knowledge to prioritize research [88]. Strategies to overcome this phenomenon might further improve the performance of the co-scientist as a tool for scientific discovery.

Improved multimodal reasoning and tool-use capabilities. Some of the most interesting data in scientific publications is not written in text but may be encoded visually in figures and charts. However, even state-of-the-art frontier models may not comprehensively utilize such data with optimal reasoning [89] and the AI co-scientist system is unlikely to be an exception. Stronger benchmarks and evaluations are necessary to improve these capabilities. We have also not evaluated the ability of our system to reason over and integrate information from domain-specific biomedical multimodal datasets (such as large multi-omics datasets) and knowledge graphs. More work is needed to integrate the AI co-scientist system with specialized scientific tools, AI models and databases, and evaluate the ability to utilize them effectively.

Inherited limitations of frontier LLMs. LLM limitations include imperfect factuality and hallucinations, which may be propagated in the co-scientist system. The system's reliance on existing LLMs and web-search, while providing immediate access to broad knowledge, may propagate errors of factuality, biases or limitations present in those resources.

Need for better metrics and broader evaluations. While the current AI co-scientist evaluation includes AI auto-ratings, expert reviews and targeted *in vitro* validations, the evaluation of system performance remains preliminary. A comprehensive, systematic evaluation across diverse biomedical and scientific disciplines is necessary to determine the generalizability of co-scientist. Furthermore, the system requires continued improvement to produce outputs that meet the rigor and detail of high-quality publications. Furthermore, the Elo rating implemented to help the system self-improve for hypothesis generation is a limited auto-evaluation metric. Continued investigation into alternative, more objective, less intrinsically-favored, evaluation metrics that better represent perspectives and preferences from expert scientists could strengthen future work.

Limitations of existing validations. At present, the AI co-scientist focuses on identifying potential therapeutic targets and mechanisms, but many not be addressing the complexities of drug delivery systems. Pharmaceutical factors such as tissue-specific targeting, formulation requirements, and delivery efficiency—while critical for clinical translation—remain beyond the scope of the present system.

The AI co-scientist is currently also not designed to generate comprehensive clinical trial designs or to fully account for factors such as drug bioavailability, pharmacokinetics, and any complex drug interactions when applied for drug repurposing or discovery. These aspects require much deeper understanding, extensive expertise, and appropriate data beyond the scope of the current system. A dedicated translational scientific team is needed for onward clinical translation of the predictions. These limitations also highlights the need for continued development and integration of the system with more tools, such as specialized AI models and real-time databases.

6 Safety and Ethical Implications

While AI systems such as the co-scientist offers the potential to accelerate scientific discovery, it also poses significant safety and ethical challenges, distinct from its impact on the scientific method itself. Safety risks center on the dual-use and the possibility that scientific breakthroughs could be exploited for harmful purposes. Ethical risks, conversely, involve research that contradicts established ethical norms and conventions within specific scientific disciplines. We review these distinct risk categories, emphasizing that further research is crucial to fully understand and mitigate them.

Evolving ethics frameworks, policy and regulations for advanced AI use in scientific endeavors. Research ethics is a central aspect of scientific endeavor and a prominent research field in its own right [90–95]. A key focus is directing research towards positive societal impact, although questions remain about potentially dual-use knowledge [96–100].

Core ethics principles are being complemented by emerging regulation, and formal processes involving organizational ethics reviews that are meant to provide an assessment of adherence to the code of conduct, as well as an assessment of present and future risks involved with research proposals [101–104].

The acceleration of science through AI, especially with advanced agentic AI systems, requires advances in science and AI ethics policy and regulation [105, 106]. This adaptation is crucial to address the changing research landscape and the unique risks associated with AI agents of varying capabilities and autonomy.

Advancements in AI systems, like the co-scientist, require moving beyond the limited ethical considerations designed for earlier, specialized AI models with restricted application and action spaces [107]. Some preliminary frameworks have developed to understand the impact of LLM agents in science, specifically mapping risks across user intent, domain, and broader impact [108].

Dual-use risks and technical safeguards. Beyond the scientific domain, broad frameworks are being developed for evaluating the emergence of potentially dangerous capabilities in AI agents [109–111]. These frameworks assess capabilities related to persuasion, deception, cybersecurity, self-proliferation, and self-reasoning. As AI agents advance, safety evaluations in science must integrate these broader assessments. A long-term risk is that agentic systems could develop intrinsic goals influencing research directions. Human susceptibility to AI manipulation, already observed in other contexts [112], underscores the need for robust frameworks ensuring instruction-following and value alignment.

More immediately on a shorter time-scale, technical safeguards are needed to address unethical research queries, malicious user intent, and the potential for extracting dangerous or dual-use knowledge from scientific AI systems. Because verification is computationally 'easier' than generation, significant research focuses on using advanced LLMs as 'critics' or 'judges' to evaluate both user queries and AI outputs acting as a scalable oversight mechanism. These critics operate based on predefined criteria, provided through direct instructions, examples (few-shot or many-shot prompting), or fine-tuning [113–118]. They can also leverage external tools for grounding [119] and have shown promise in multimodal scenarios [120]. However, limitations remain; human expert involvement is crucial, as LLMs may not align with human judgment in specialized domains [121].

Adversarial robustness of scientific AI systems. Recognizing and mitigating adversarial attacks is a crucial, ongoing research area in the development of foundation models and advanced AI assistants [85, 122–128]. While manual red teaming has identified vulnerabilities, automated approaches now allow for optimizing prompt suffixes to bypass safety measures, using techniques like greedy, gradient-based, or evolutionary methods [129, 130]. Attacks can also exploit few-shot demonstrations, in-context learning [131, 132], and multimodal inputs [133]. Furthermore, LLMs can be used to generate and refine attacks against other LLMs [134], and attacks can be iterative, spanning multiple steps [135]. Defenses are being developed to counter both human and automated attacks, which is increasingly important in an agentic AI future [136].

Advances in post-training of base models will likely improve overall adversarial robustness. However, domain-specific recognition of malicious use may still require dedicated development and integration into scientific AI assistants. In AI systems employing iterative reasoning (e.g., request interpretation, hypothesis generation, internal thoughts, evaluation, user queries), each component must be tested independently. This comprehensive testing should account for all potential failure modes, including the handling of unsafe queries, the safety of hypotheses (intermediate and final), and the accuracy of internal checks and filters.

Need for a comprehensive safety approach. Scientific AI assistants, like the co-scientist, require integrated, configurable guidelines within their safeguards. Developers should anticipate the complexity of this challenge and prioritize flexible safeguarding to rapidly incorporate community feedback. These semantic safeguards may need to be augmented by traditional software safety measures, including trusted testers, gradual feature rollouts, access controls, request logging, and flagging uncertain outputs for manual review.

Ensuring the safety of these systems, in line with existing AI safety guidelines [137, 138], necessitates a multi-pronged approach. This includes:

- Comprehensive threat modeling to identify potential vulnerabilities.
- Defense mechanisms for each identified threat.
- Extensive red-teaming and security testing.
- Rapid response procedures for quick resolution of issues including vulnerability patches.
- Continuous monitoring and performance tracking.

These considerations highlight the need for responsible development, governance and careful deployment of technologies designed for advancing science, appropriate safeguards and ethical guidelines and close compliance with applicable regulations. They also further underscore the need for broad community engagement and an inclusive development of best practices and recommendations around safe and ethical use for AI in science.

Current safeguards in the AI co-scientist. To mitigate these risks, the AI co-scientist currently employs the following safety mechanisms:

- **Reliance on public frontier LLMs.** The system utilizes established public Gemini 2.0 models, which already incorporate extensive safety evaluation and safeguards.
- **Initial research goal safety review.** Upon input, each research goal undergoes automated safety evaluation. Goals deemed potentially unsafe are rejected.
- **Research hypothesis safety review.** Generated hypotheses are reviewed for safety, even when the overarching research goal is deemed safe. Potentially unsafe hypotheses are excluded from the tournament, not developed any further, and are not presented to the user.

- **Continuous monitoring of research directions.** A meta-review agent provides an overview of research directions, enabling the AI co-scientist to continuously monitor for potential safety concerns and alert users if a research direction is detected as being potentially unsafe.
- **Explainability and transparency.** All system components, including the safety review, provide not only the final recommendation but also a detailed reasoning trace that can be used to justify and audit system decisions.
- **Comprehensive logging.** All system activities are logged and stored for future analysis and auditing.
- **Safety evaluations and red teaming.** A preliminary red teaming effort has been undertaken to ensure that the current implementation of unsafe research goal detection is robust and accurate. This evaluation includes an assessment of the system behavior when presented with 1,200 adversarial research goals across 40 distinct topic areas as discussed in Section 4.4.
- **Trusted tester program.** We are enthused by the early promise of the AI co-scientist system and believe it is important to more rigorously understand its strengths and limitations in many more areas of science and biomedicine; alongside making the system available to many more researchers who it is intended to support and assist. To facilitate this responsibly and with rigour, we will be enabling access to the system for scientists through a Trusted Tester Program to gather real-world feedback on the utility and robustness of the system.

Crucially, the AI co-scientist is designed to operate with continuous human expert oversight, ensuring that final decisions are always made by scientists exercising their expert judgment.

7 Future Work

Immediate improvements. The AI co-scientist is in its early development, with many opportunities for improvement. Immediate improvement opportunities include enhanced literature reviews, cross-checks with external tools, improved factuality checking, and increased citation recall to minimize missed relevant research. Coherence checks would also improve the system by reducing the burden of reviewing flawed hypotheses.

Expanded evaluations. Developing more objective evaluation metrics, potentially incorporating automated literature-based validation and simulated experiments, is a key area. Methods to mitigate biases or error patterns inherited from the base LLMs are also important, alongside analysis of the complementarity and optimal combination of different agent components.

A critical need is a larger-scale evaluation involving more subject matter experts with diverse, high-resolution research goals. Stress-testing the system at every level of resolution (from disease mechanisms to protein design, and expanding to other scientific disciplines) will reveal further areas for improvement. Finally, since laboratory resources are limited, improved evaluation frameworks could assist with hypothesis selection.

Capabilities advancements. Several opportunities remain to expand co-scientist's capabilities. Reinforcement learning could enhance hypothesis ranking, proposal generation, and evolutionary refinement.

Currently, the system assesses text from open-access publications but not images, data sets, or major public databases. Integrating these publicly available resources would significantly enhance the co-scientist's ability to generate and justify proposed hypotheses.

Future work will focus on handling more complex experimental designs, such as multi-step experiments and those involving conditional logic. Integrating co-scientist with laboratory automation systems could potentially create a closed-loop for validation and a grounded basis for iterative improvement. Exploring more structured user interfaces for providing feedback and insights from targeted user research studies, beyond free text, could improve the efficiency of human-AI collaboration in this paradigm.

8 Discussion

Our study represents an initial foray into accelerating novel scientific discovery with agentic AI systems and here, we discuss some of the broader implications. The co-scientist iteratively refines its generated hypotheses through a generate, debate, evolve" approach with specialized agents under the hood. This design creates a self-improving cycle for research hypothesis generation, as measured by automated evaluation metrics, and

showcases the potential benefits of test-time compute scaling for scientific reasoning.

Instead of brute-force generation of a vast number of hypotheses and relying on volume to chance into a few potentially useful ones, the system is designed to mimic key aspects of the scientific reasoning method in an intelligent manner. As detailed in Section 3, the co-scientist employs principled internal mechanisms, including scientific debates, tournaments, iterative refinement, and human feedback loops to progressively improve the quality of its proposals, and converge on high quality and well-reasoned hypotheses.

AI co-scientist novelty is grounded in prior evidence. The AI co-scientist facilitates the generation of novel scientific hypotheses and uncovering new insights by synthesizing extensive literature and identifying latent relationships. While its primary utility in its current form may lie in enabling more incremental advancements — such as the computational repurposing of existing therapeutics — it may also be able to support exploratory, breakthrough research. When researchers define such open-ended research goals requiring complex and out-of-the box thinking, the system may produce outputs of varying confidence. Therefore, rigorous validation and critical appraisal by domain experts remain paramount. This system is intended to augment, not supplant, human scientific reasoning, empowering researchers to accelerate discovery while maintaining intellectual control over the generated insights. We further expand on the novelty aspects in the specific context of the applications considered in this work in Section A.1.

Multiple experimental validations of novel co-scientist hypotheses. Notably, this work demonstrates the validation of co-scientist hypotheses via experimental findings in multiple laboratories. In drug repurposing, co-scientist identifies novel candidates for AML that demonstrated *in vitro* efficacy at clinically relevant concentrations, including the identification of new repurposing opportunities beyond current preclinical knowledge. For liver fibrosis, the co-scientist proposes new epigenetic treatment targets, with subsequent *in vitro* experiments validating the anti-fibrotic activity of several suggested compounds, including an FDA-approved drug. In the realm of antimicrobial resistance, the co-scientist independently recapitulates a novel, unpublished finding regarding the mechanism of cf-PICI transfer between bacterial species. Early results over several queries of varying scientific complexity suggests the co-scientist has a potential to contribute to discovery within various biomedical domains.

Test-time compute scaling with scientific reasoning priors and inductive biases. In the experiments reported here, the co-scientist did not require specialized pre-training, post-training, or a reinforcement learning framework. It leverages the capabilities of existing base LLMs, potentially benefiting from updates to those models without requiring retraining of the co-scientist system itself, which presents advantages of compute efficiency and generalizability. The system's architecture incorporates self-play, internal consistency checks, and tournament-based ranking, which support iterative hypothesis generation, evaluation, and refinement. This is reflected in the observed improvement in hypothesis quality over time. This self-evolution can be improved further by expanded tool use integration, including database queries, allowing the co-scientist to ground its proposals in existing knowledge and identify novel connections. In the future, we may leverage data and tournament ranking generated by the co-scientist itself as feedback to improve the whole system using reinforcement learning.

Frontier LLM advancements and the AI co-scientist. The frontier LLMs used within the co-scientist system have demonstrated a continuing trend of rapidly improved capabilities, including reasoning, logic, and also some aspects of scientific literature comprehension. As our system is designed to be model-agnostic, we hypothesize that further improvements in frontier LLMs will also result in improved co-scientist performance, and enable new avenues of research including optimal agentic use of tools.

Implications for drug repurposing and discovery. These advancements have significant implications for various biomedical and scientific domains. For example, the integration of the co-scientist into the drug candidate selection process represents a significant advancement in evidence-based drug repurposing. Beyond simple literature mining, the co-scientist maybe capable of synthesizing novel mechanistic insights by connecting molecular pathways, existing preclinical evidence, and potential therapeutic applications into structured, testable specific-aims. This capability is particularly valuable as it provides researchers with literature-supported rationales and suggests specific experimental approaches for validation. Notably, the

co-scientist's structured output can be leveraged to develop comprehensive single-patient IND (Investigational New Drug) applications for compassionate use cases. By systematically presenting mechanistic evidence, relevant preclinical data, and proposed monitoring parameters, the co-scientist facilitates the development of well-reasoned treatment protocols for patients with refractory (treatment-resistant) disease who have exhausted standard therapeutic options and are ineligible for clinical trials. This application is particularly valuable in rare or aggressive diseases where traditional drug development timelines may not align with urgent patient needs. The platform's ability to rapidly generate evidence-based therapeutic hypotheses, complete with safety considerations and monitoring parameters, can help clinicians and regulatory bodies make informed decisions about compassionate use applications while maintaining scientific rigor.

The application of the co-scientist in drug repurposing presents a very compelling opportunity for orphan drugs, where extensive safety and clinical data already exist from their original rare disease indications. Given that Phase III clinical trials can cost hundreds of millions of dollars, repurposing these well-characterized therapeutics offers an efficient path to expanding treatment options across multiple diseases. This is especially relevant as orphan drugs often target fundamental biological pathways that may be relevant in other conditions, but these connections might not be immediately apparent through traditional research approaches. By systematically evaluating existing clinical data, safety outcomes, and mechanistic insights, the co-scientist can help identify promising new therapeutic applications while taking advantage of the investment already made in drug development and safety validation. This approach not only maximizes the utility of existing therapeutics but also provides a more rapid path to addressing unmet medical needs across a broader patient population.

More broadly, the co-scientist may also be potentially impactful throughout the entire drug discovery spectrum as evidenced by the early work on co-scientist assisted target discovery for liver fibrosis.

Automation bias and impact on human scientific creativity. Realizing the full potential of AI in biomedicine and science requires proactively addressing potential pitfalls. Over-reliance on AI-generated suggestions in collaborative AI systems could diminish critical thinking and increase homogeneity in research. Studies on AI's impact on creativity and ideation show mixed results; some suggest a risk of homogenization of ideas across populations [139], while others are less conclusive [140]. The correlated success / failure modes of LLMs [141], due to similar training data, could also artificially narrow scientific inquiry. Furthermore, AI system blind spots and performance variations across research domains must be considered. Therefore, scalable factuality and verification methods, alongside peer review and careful consideration of potential biases, are essential. Careful design and use of systems like the co-scientist are crucial to mitigate these risks.

AI as a catalyst for both scientific discovery and equity. Despite these risks, AI holds immense potential to democratize access to scientific information and accelerate discovery, particularly benefiting historically neglected and resource-constrained areas [142, 143]. In essence, AI can "raise the tide" of scientific progress, lifting all boats, especially those that have historically been left behind. Realizing this potential requires strategic investments and careful calibration of AI systems to foster ideation and innovation while minimizing false positives. This includes focusing on historically neglected research topics and addressing variations in performance across different scientific domains with varying amounts of pre-existing data. While current AI systems may tend to produce incremental ideas and research hypotheses, ongoing development aims to create systems capable of generating truly original, unorthodox and transformative scientific theories. Proactive mitigation of these challenges will ensure that AI serves as a powerful tool for all scientists, promoting a more equitable and innovative future for scientific explorations.

9 Conclusion

The AI co-scientist represents a promising step towards AI-assisted augmentation of scientists and acceleration of scientific discovery. Its ability to generate novel 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 the AI co-scientist system 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.

Acknowledgments

We thank our teammates Subhashini Venugopalan, John Platt, Erica Brand, and Yun Liu for their detailed technical feedback on the manuscript. We thank Jakob T Rostoel, Cora Chmielowska and Jonasz B Patkowski from Imperial College London and Jakkapong Inchai, Weida Liu, and Wenlong Ren from Stanford University for providing expert feedback on the AI system introduced in this work, and the lab of Ravi Majeti from Stanford University for generously providing the AML cell lines used in this work. We thank Ritu Raman, Ryan Flynn, Charlie Hempstead, Lord Ara Darzi, Omar Abudayyeh, Jonathan Gootenberg, Nic Fishman, Jason Lequyer, Dan Leesman, Ravi Solanki, Dennis Gong and Ananthan Sadagopan for feedback on different aspects of the AI system and the work. We also thank Maen Abdelrahim, Ethan Burns, Preethi Prasad and Hanh Mai for their clinical expertise and expert evaluation.

We thank our teammates Thomas Wagner, Alessio Orlandi, Natasha Latysheva, Nir Kerem, Yaniv Carmel, Hussein Hassan Harrirou, Laurynas Tamulevičius and Grzegorz Glowaty for their technical support. We thank Taylor Goddu, Resham Parikh, Siyi Kou, Rachelle Sico, Amanda Ferber, Cat Kozlowski, Alison Lentz, KK Walker, Roma Ruparel, Jenn Sturgeon, Lauren Winer, Juanita Bawagan, Ed-Allt Graham, Tori Milner, MK Blake, Jack Mason, Erika Radhansson, Indranil Ghosh, Jay Nayar, Brian Cappy, Celeste Grade, Abi Jones, Laura Vardoulakis, Lizzie Dorfman, Ashmi Chakraborthy, Delia Williams-Falokun, Maggie Shiels, Kalyan Pamarthy, Sarah Brown, Christian Wright, and S. Sara Mahdavi for their support and guidance during the course of this project. Finally, we thank Michael Brenner, Zoubin Ghahramani, Dale Webster, Joelle Barral, Michael Howell, Susan Thomas, Karen DeSalvo, Jason Freidenfelds, Ronit Levavi Morad, Vladimir Vuskovic, Ali Eslami, Anna Koivuniemi, Greg Corrado, Royal Hansen, Andy Berndt, Noam Shazeer, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean and Demis Hassabis for their support of this work.