Cell
Leading Edge

Perspective

Empowering biomedical discovery with AI agents

Shanghua Gao,[1] Ada Fang,[1,2,8,11] Yepeng Huang,[1,3,11] Valentina Giunchiglia,[1,4,11] Ayush Noori,[1,5,11] Jonathan Richard Schwarz,[1] Yasha Ektefaie,[1,6] Jovana Kondic,[7] and Marinka Zitnik[1,8,9,10,*]
[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA
[3]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA
[4]Department of Brain Sciences, Imperial College London, London, UK
[5]Harvard College, Cambridge, MA, USA
[6]Program in Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[7]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA
[8]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Allston, MA, USA
[9]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[10]Harvard Data Science Initiative, Cambridge, MA, USA
[11]These authors contributed equally
*Correspondence: marinka@hms.harvard.edu
https://doi.org/10.1016/j.cell.2024.09.022

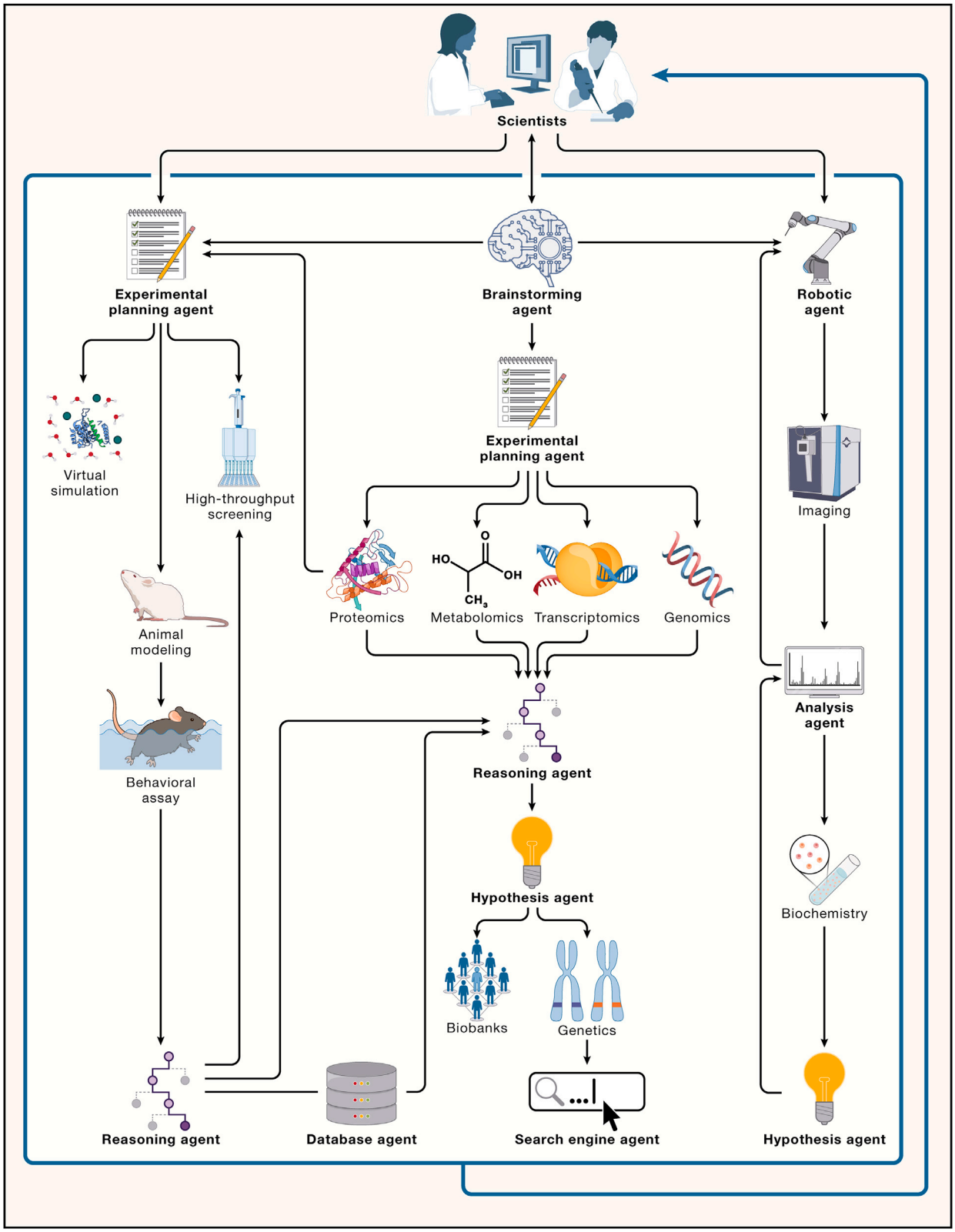
SUMMARY

We envision "AI scientists" as systems capable of skeptical learning and reasoning that empower biomedical research through collaborative agents that integrate AI models and biomedical tools with experimental platforms. Rather than taking humans out of the discovery process, biomedical AI agents combine human creativity and expertise with AI's ability to analyze large datasets, navigate hypothesis spaces, and execute repetitive tasks. AI agents are poised to be proficient in various tasks, planning discovery workflows and performing self-assessment to identify and mitigate gaps in their knowledge. These agents use large language models and generative models to feature structured memory for continual learning and use machine learning tools to incorporate scientific knowledge, biological principles, and theories. AI agents can impact areas ranging from virtual cell simulation, programmable control of phenotypes, and the design of cellular circuits to developing new therapies.

INTRODUCTION

A long-standing ambition for artificial intelligence (AI) is the development of AI systems that can make major scientific discoveries, learn on their own, and acquire knowledge autonomously. While this concept of an "AI scientist" is aspirational, advances in agent-based AI pave the way to the development of *AI agents* as conversable systems capable of reflective learning and reasoning that coordinate large language models (LLMs), machine learning (ML) tools, experimental platforms, or even combinations of them[1–4] (Figure 1). The complexity of biology calls for approaches that flexibly decompose complex problems into actionable tasks. AI agents can break down a problem into manageable subtasks, which can then be addressed by agents with specialized functions for targeted problem solving and integration of scientific knowledge.[1,5] In the near future, AI agents can accelerate discovery workflows by making them faster and more resource-efficient. AI agents improve the efficiency of routine tasks, automate repetitive processes, and analyze large datasets to navigate hypothesis spaces at a scale and precision that surpasses current human-driven efforts. This automation allows for continuous, high-throughput research that would be impossible for human researchers to perform alone at the same scale or speed. Looking ahead, AI agents can provide insights beyond what traditional machine learning alone can achieve by making predictions across temporal and spatial scales before experimental measurements at those scales are available. Ultimately, they may help uncover new modes of behavior within biological systems.[5]

This vision is possible thanks to advances in LLMs,[6–8] multimodal learning, and generative models. Chat-optimized LLMs, such as GPT-4,[9] can incorporate feedback, enabling AI agents to cooperate through conversations with each other and with humans.[10] These conversations can involve agents seeking human feedback and critique and identifying gaps in their knowledge.[11,12] Then, since a single LLM can exhibit a broad range of capabilities—especially when configured with appropriate prompts and inference settings—conversations between differently configured agents can combine these capabilities in a modular manner.[13] LLMs have also demonstrated the ability to solve complex tasks by breaking them into subtasks.[14,15] However, suppose we follow conventional approaches to foundation models such as LLMs and other large pre-trained models. In that case, we may not develop AI agents that can generate novel hypotheses because such novelty would not have been in the data used to train the model, suggesting that current foundation



(legend on next page)

models alone are not sufficient for AI scientists. Using LLMs as a comparison, generating novel hypotheses requires creativity and grounding in scientific knowledge, whereas generating novel text requires adherence to semantic and syntactic rules,[16] so the latter approach aligns well with techniques for next-token prediction within LLMs, while the former might not.

Here, we offer a perspective that AI scientists can be realized as AI agents backed by humans, LLMs, ML models, and other tools like experimental platforms that together form compound AI systems. An AI agent should be able to formulate biomedical hypotheses, critically evaluate them, characterize their uncertainty, and use that as a driver to acquire and refine its scientific knowledge bases in a way that human scientists can trust.[17] AI agents should be designed to adapt to new biological insights, incorporate the latest scientific findings, and refine hypotheses based on experimental results. This adaptability ensures agents remain relevant in the face of rapidly evolving biological data,[16] balancing between encoding new findings and retaining old knowledge.[18]

Realizing this perspective shift, biomedical AI agents can impact areas ranging from virtual cell simulation, programmable control of phenotypes, and the design of cellular circuits to developing new therapies. Virtual cell simulation involves creating detailed models of cellular processes where AI can predict the effects of genetic modifications or drug treatments on cell behavior. This can allow for an understanding of cellular mechanisms and generation of testable hypotheses, reducing the time and cost of traditional methods. Programmable control of phenotypes leverages AI agents to design precise genetic modifications to study gene functions. For example, CRISPR-based gene editing guided by an AI agent can activate or inhibit specific genes across large cell populations in a multi-round editing campaign. Each round involves identifying the next edit based on the user-specified target phenotype and experimental readout from the previous round. Designing cellular circuits involves using AI agents to predict the behavior of genetic components and optimize their arrangement to create circuits that perform tasks such as sensing environmental changes or producing therapeutic proteins.

Ethical considerations arise from biomedical AI agents.[19,20] Allowing them to make changes in environments through ML tools or calls to experimental platforms can be dangerous. Safeguards need to be in place to prevent harm.[21] Conversely, discovery workflows might include conversations between AI agents (but no interaction with environments is allowed). In that case, we need to consider the impact of such interactions on human scientists and their reliance on AI agents. Additionally, a key challenge specific to biomedical AI agents is the lack of large, diverse experimental datasets beyond the current focus areas in structural and cell biology. AI agents must represent biomedical knowledge efficiently, generalize well to new tasks, and acquire new skills with minimal or no additional training. While AI agents can empower research and support operations under human oversight, their potential impact and associated challenges underscore the importance of responsible implementation.

EVOLVING USE OF DATA-DRIVEN MODELS IN BIOMEDICAL RESEARCH

Over the past several decades, data-driven models have reshaped biomedical research by developing databases (DBs), search engines, ML, and interactive and foundation learning models (Figure 2). These models have advanced modeling of proteins,[22–26] genes,[27] phenotypes,[28] clinical outcomes,[29–31] and chemical compounds[32,33] through mining of biomedical data.

DBs and search engines

In biological research, DBs[34–36] aggregate knowledge from experiments and studies, offering searchable repositories containing standardized biological data vocabularies. An example of such a DB is the AlphaFold Protein Structure DB,[37] which includes more than 200 million protein structures predicted by AlphaFold.[38] Molecular search engines retrieve information from these DBs.[39–41] FoldSeek[42] retrieves protein structures from the AlphaFold DB by translating query structures into 3D interaction alphabet sequences and using pre-trained substitution matrices. Search engines are designed to retrieve information based on specific queries, lacking the ability to refine these queries through reasoning. They cannot iteratively process obtained information to refine results or customize subsequent actions. Additionally, while DBs reduce the risk of misinformation through curated data, they lack mechanisms to identify and remove irrelevant information.

Distinct from search engines, AI agents are capable of reasoning to formulate search queries and subsequently acquire information. Curated DBs offer structured and factual information, aiding in reducing the risks associated with misinformation potentially generated by agent hallucinations.[43,44] For example, the retrieval-augmented generation (RAG)[44] is equipped for AI agents to answer questions based on scientific literature. A notable feature of these agents is their ability to retrieve information when needed and to create and iteratively process the obtained passages. This reflection process makes the agent controllable during inference, allowing for customization of its actions to meet task requirements beyond what is possible using search engines and DB queries.

Figure 1. Empowering biomedical research with AI agents

AI agents are laying the groundwork for AI scientists as compound AI systems capable of skeptical learning and reasoning. These multi-agent systems consist of agents based on conversable large language models (LLMs) and can coordinate machine learning (ML) tools, experimental platforms, humans, or even combinations of them. Robotic agent, AI agent that operates robotic hardware for physical experiments; database agent, AI agent that can access information in databases via function calling and application programming interfaces (APIs); reasoning agent, AI agent capable of direct reasoning and reasoning with feedback; hypothesis agent, AI agent that is creative and reflective when developing hypotheses, capable of characterizing its own uncertainty and using that as a driver to refine its scientific knowledge bases; brainstorming agent, AI agent that generates a broad spectrum of research ideas; search engine agent, AI agent that uses search engines as tools to rapidly gather information; analysis agent, AI agent capable of analyzing experimental results to summarize findings and synthesize concepts; and experimental planning agent, AI agent that optimizes an experimental protocol for execution.

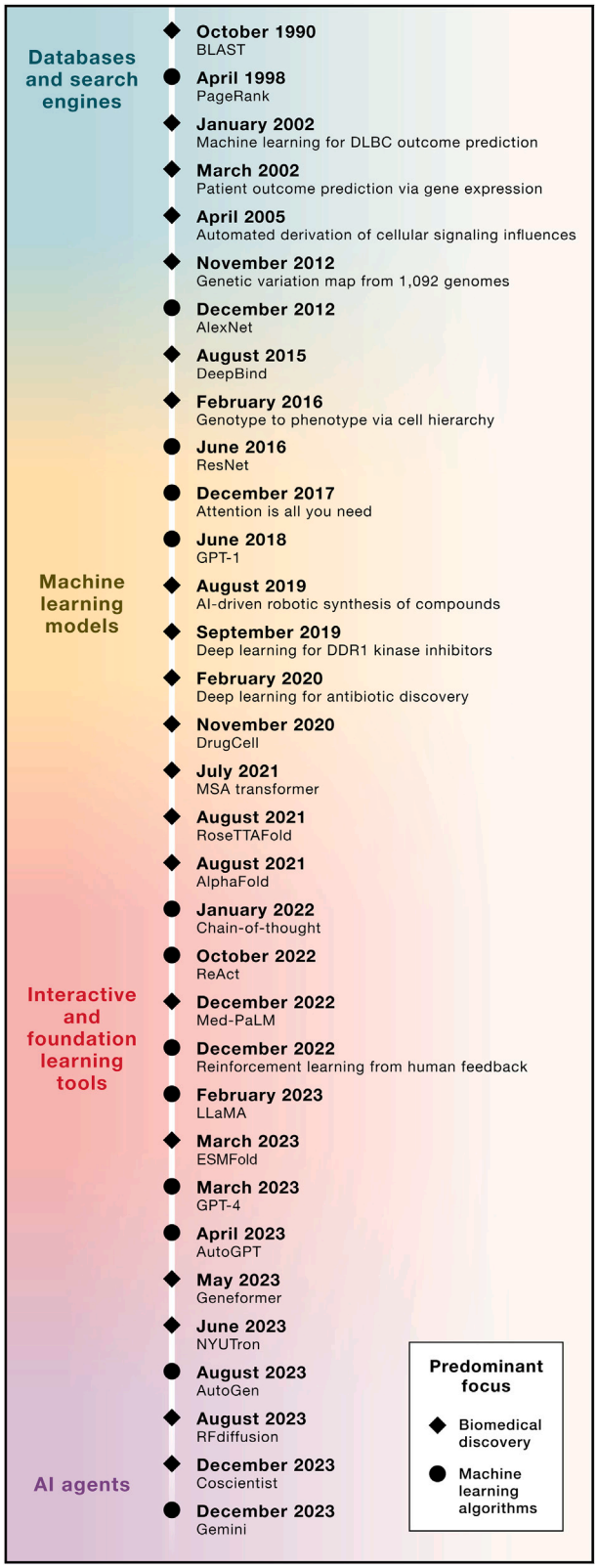


Figure 2. Evolving use of data-driven models

Data-driven approaches, from databases and search engines, ML, and interactive learning models to advanced agent systems, have reshaped biomedical research throughout the last several decades. Circles represent studies focused predominantly on algorithmic ML innovation; diamonds are used to indicate representative studies that used AI for biomedical discovery.

ML models

Beyond information retrieval, ML models excel in identifying patterns and assimilating latent knowledge to generalize predictions about novel data.[45,46] Existing ML models typically require specialized models for each task and do not possess the reasoning and interactive capabilities that distinguish AI agents. An example is the AlphaFold,[38] which predicts 3D protein structures with high accuracy using multi-sequence alignment with a deep learning model but is tailored for protein folding. AI agents represent an evolution in ML models, building on the foundations of successes such as the transformer architecture[47] and generative pretraining.[8] These agents' reasoning and interactive capabilities distinguish them from ML models, which typically require specialized models for each task. Unlike traditional ML models, agents assess the evolving environment, which is valuable for modeling dynamic biological systems.

Interactive learning models

Interactive learning, often referred to as active learning[48] and reinforcement learning,[49] represents a further advancement in ML models by incorporating exploration mechanisms and human feedback. Active learning strategies can help build models for datasets with small sample sizes when conventional ML models might be insufficient due to limited statistical power. It selectively queries the most informative data points for labeling and optimizing the learning process, which improves how models learn with data. Reinforcement learning involves an agent learning how to act by observing the results of past actions in an environment, mirroring the trial-and-error approach. In biological research, interactive learning has been used for small molecule design,[50] protein design,[51,52] drug discovery,[53,54] perturbation experiment design,[55] and cancer screening.[56] For instance, GENTRL[50] uses reinforcement learning to navigate the chemical space and identify chemical compounds that can act against biological targets. However, interactive models are predominantly designed for narrow use cases and struggle to generalize to new goals without retraining the models from scratch. Leveraging interactive learning, AI agents achieve greater autonomy in information retrieval tasks. Active learning improves training efficiency through data labeling selected to maximize model performance. However, AI agents extend beyond this data-centric approach; for example, reinforcement learning with human feedback (RLHF)[49] uses a "reward model" to train an LLM-based agent with direct human feedback to understand human instruction naturally.

AI agents

Biomedical AI agents have advanced capabilities, including proactive information acquisition through perception modules, interaction with tools, reasoning, and engaging with and learning from their environments. Agents use external tools, such as lab equipment, and have perception modules, such as integrated visual ML tools, to receive information from the environment. Agents can incorporate search engines and ML tools and

process information across data modalities via perception modules to generate hypotheses and refine them based on scientific evidence.[1,2]

TYPES OF BIOMEDICAL AI AGENTS

The prevailing approach to building agents is to use LLMs, where a single LLM is programmed to perform various roles. However, beyond LLM agents, we envision multi-agent systems for discovery workflows that combine heterogeneous agents (Figure 1) consisting of ML tools, domain-specific specialized tools, and human experts. Given that much of biomedical research is not text-based, such agents have broader applicability to biomedicine than LLM-based agents alone.

LLM-based AI agents

Programming a single LLM with diverse roles equips LLM-based agents with conversational interfaces that emulate human expertise and can access tools[57,58] (Figure 3A). The rationale behind this approach stems from pretraining an LLM to encode general knowledge, followed by in-domain fine-tuning of the LLM to encode domain-specific specialist knowledge, and aligning the LLM with human users through role-playing and conversation. Instruction tuning[59] can be used for the former by training the LLM to follow human instruction through prompt examples, including dialogues that incorporate biological reasoning.[60] Additionally, RLHF optimizes LLM performance by selecting the most human-preferred outputs from a range of responses to specific prompts, further aligning LLMs with human roles. Consequently, a single LLM, programmed to fulfill multiple roles, can provide a more practical and effective solution than developing specialized models. By assigning specific roles, the agents can replicate the specialized knowledge of experts across various fields, such as structural biology, genetics, and chemistry, surpassing the capabilities of querying a non-specialized LLM[61] and performing tasks previously not possible.[62] Early results in clinical medicine question-answering suggest that assigning specific roles, such as clinicians, to GPT-4[61] can achieve better performance in terms of accuracy on multiple-choice benchmarks compared with using domain-specialized LLMs like BioGPT,[63] NYUTron,[64] and Med-PaLM.[65,66]

We envision three approaches for assigning roles to biological AI agents: domain-specific fine-tuning, in-context learning, and automatic generation of agentic roles. The first approach involves instruction tuning an LLM across many biological tasks to ground the LLM in the biological domain, followed by RLHF to ensure that the tuned LLM performs tasks aligned with scientists' goals and needs. The second approach uses in-context learning of LLMs[67] to process longer contextual information provided in inputs, such as biologist-generated instructions, enabling agents to grasp the domain context for each task. This approach is supported by using textual prompts to define agent roles.[62,68] Both strategies require biologists to gather task-specific data or craft precise prompts. However, since human-defined roles might not always guide agents as intended, there is a growing shift toward granting LLM-based agents greater autonomy in defining their roles. This shift in role definition enables agents to autonomously generate and refine role prompts and engage in self-directed learning and role identification. For instance, an agent's ability to evolve and tailor its prompts in reaction to user inputs has been demonstrated in Fernando et al.[69] Additionally, self-referential learning frameworks can be employed to optimize prompt design when assigning roles to agents,[70] enabling them to transition from task executors to entities capable of autonomous setup.

The agent system, comprising a single LLM prompted to adopt various roles, has shown to be a valuable support tool in scientific research. Studies suggest that agents set up to perform specific roles exhibit enhanced capabilities compared with either sequentially querying a single LLM or employing a single tool repetitively. A case in point is Coscientist,[1] which shows the potential of GPT-4-based agents for chemical research tasks, including optimizing reactions for palladium-catalyzed cross-couplings. Within Coscientist, GPT-4 undertakes the role of a planner, serving as a research assistant. The agent uses in-context prompts to use tools such as web and documentation search and code execution via Python application programming interface (API) and symbolic lab language (SLL).[1] To complete tasks that require access to a physical device, the planning agent starts with a prompt provided by the scientist and uses search tools to compile documentation for the experiment. Following this, the agent generates SLL code and executes it, which entails transferring it onto the device and controlling the device.

Multi-agent AI systems

LLM-based agents implemented through autoregressive LLM approaches acquire skills such as planning and reasoning by emulating observed behaviors in training datasets. However, this mimicry-based learning results in limited agent capabilities, as they do not achieve a deep understanding of these behaviors.[71] Consequently, a single agent often lacks the comprehensive skill set needed to complete complex tasks. A practical alternative is deploying a multi-agent AI system, wherein the task is segmented into more manageable subtasks. This approach allows individual agents to address specific subtasks efficiently, even with incomplete capabilities. Distinct from single-LLM-based agents, a multi-agent system incorporates several agents endowed with specialized capabilities, tools, and domain-specific knowledge. For successful task execution, these agents must conform to working protocols. Such cooperative efforts equip LLMs with unique roles, specialized knowledge bases, and varied toolsets, simulating an interdisciplinary team of biology specialists. This approach is akin to the diverse expertise found across departments within a university or an institute.

In the following, we introduce five collaborative designs for multi-agent systems.

Brainstorming agents

Brainstorming research ideas with multiple agents (Figure 3B) constitutes a collaborative session to generate a broad spectrum of research concepts through the joint expertise of scientists and agents. In such sessions, agents are prompted to contribute ideas, prioritizing the volume of contributions over their initial quality to foster creativity and innovation. This method encourages the proposal of unconventional and novel ideas, allowing participants to build upon the suggestions of others to uncover new avenues of inquiry while withholding judgment or critique. The process enables agents to apply their domain knowledge

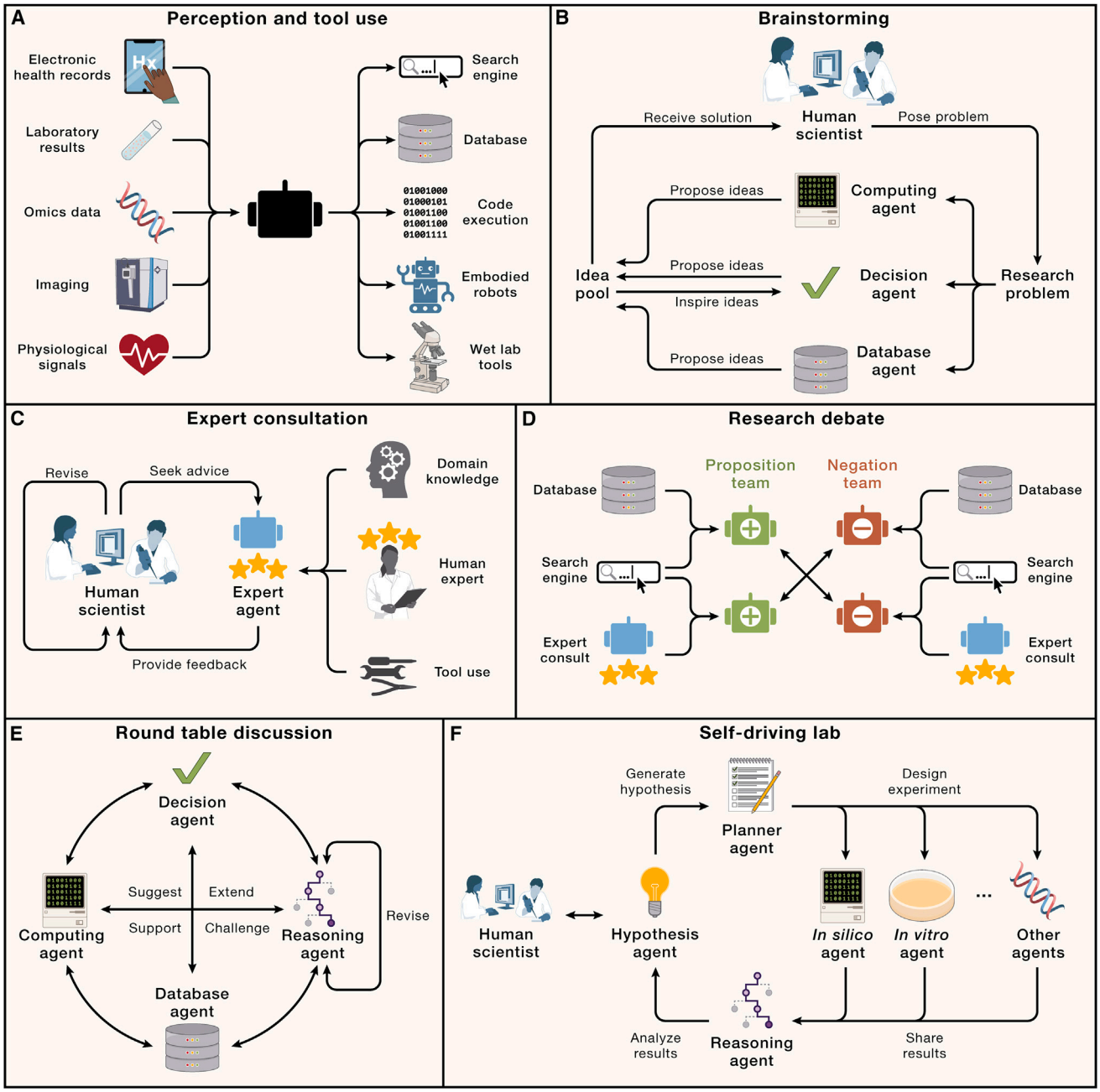


Figure 3. Diverse configurations of AI agents in biomedicine—from an LLM-based AI agent to a multi-agent system with AI models, tools, and integrated physical devices

(A) By programming an LLM with the role, one LLM-based agent, equipped with memory and reasoning abilities, performs multimodal perception and utilizes a range of tools, e.g., web lab tools, to accomplish specified tasks.

(B–E) Leveraging AI agents equipped with diverse roles, perception modules, tools, and domain knowledge enables collaboration between agents and scientists. This collaboration can adopt various configurations, such as expert consultation, debate, brainstorming, and roundtable discussions.

(F) Multi-agent systems can establish a self-driving laboratory wherein numerous agents collaborate on multiple iterations of biological research assisted by humans. Each cycle of research encompasses the generation of hypotheses, the design of experiments, the execution of experiments both *in silico* and *in vitro*, and the analysis of results.

Computing agent, AI agent that utilizes computational models as tools; decision agent, AI agent that makes decisions in response to given conditions; database agent, AI agent that retrieves relevant information from databases; reasoning agent, AI agent capable of direct reasoning and reasoning with feedback; expert agent, AI agent that provides professional consultation based on reliable sources, such as domain expertise, feedback from human experts, and the results of specific tools; hypothesis agent, AI agent capable of reflective learning and reasoning to generate hypotheses; planner agent, AI agent that devises plans for future actions; and *in silico/vitro* agent, AI agent that uses tools *in silico* or *in vitro* environment.

and resources to form a collective idea pool. Each agent would provide insights and generate hypotheses based on their specialized knowledge, which the group can then integrate and refine. For example, in a multi-agent system designed for Alzheimer's research, agents could specialize in microglia biology, neuronal degeneration, and neuroinflammation. To explore new therapeutic targets for Alzheimer's disease, an agent specialized in microglia biology might propose investigating the role of microglial cells in synaptic pruning, while another agent focused on neuronal degeneration could suggest examining the protective effects of certain neurotrophic factors. These diverse ideas are pooled together, allowing researchers to explore a comprehensive range of potential research directions.

Expert consultation agents

Expert consultation (Figure 3C) entails soliciting expertise from individuals or entities with specialized knowledge. This process involves expert agents gathering information from various sources and providing insights, solutions, decisions, or evaluations in response. Other agents or humans then refine their approaches based on this feedback. LLMs have the potential to assist in offering scientific critiques on research manuscripts, as demonstrated in recent studies.[72] However, LLMs lack the nuanced understanding of human reviewers and should be seen as complementary to, not a replacement for, human expertise. Similarly, an AI agent might consult another agent specialized in a specific area to refine ideas within AI systems, mirroring the mentor-mentee dynamics found in academic environments. In another example, in addressing Alzheimer's and related dementias, diagnosing Alzheimer's based on cognitive criteria might present borderline cases. Consulting an AI agent could offer additional perspectives, determining if such cases align with Alzheimer's based on brain pathology or alternative biomarkers.

Research debate agents

In a research debate (Figure 3D), two teams of agents present contrasting perspectives on a research topic, aiming to persuade the agents of the opposing team. Agents are split into two groups, each adopting distinct roles for the debate. One group gathers evidence to fortify its position using various knowledge sources and tools, while the opposing group critiques this evidence, striving to expose or neutralize its weaknesses with superior evidence. The objective for each faction is to articulate their arguments more effectively than their rivals, engaging in a systematic discourse to defend their viewpoint and challenge the veracity of their adversaries' assertions. This methodology promotes critical thinking and bolsters effective communication as each team endeavors to construct the most compelling argument supporting their stance.

Roundtable discussion agents

Roundtable discussions (Figure 3E) involve multiple agents engaging in a process that fosters the expression of diverse viewpoints to make collaborative decisions on the topics under discussion. In such sessions, agents articulate their ideas and insights, pose questions, and provide feedback on others' contributions. They then respond to these queries, refine their initial propositions based on feedback, or attempt to persuade their peers. This method promotes equal participation among all agents, urging them to contribute their expertise and perspectives, offer constructive criticism, question underlying assumptions, and suggest amendments to improve the proposed solutions. Reconcile[73] implements a multi-agent collaborative framework where multiple LLM-based agents engage in several rounds of dialogue to reach a consensus on research questions. Agents attempt to convince each other to adjust their responses and use a confidence-weighted voting mechanism to achieve a more accurate consensus than if a single LLM-based agent is used. During each discussion round, Reconcile orchestrates the interaction between agents using a "discussion prompt," which includes grouped answers and explanations produced by each agent in the preceding round, their confidence levels, and examples of human explanations for correcting answers.

Self-driving lab agents

The self-driving laboratory (Figure 3F) is a multi-agent system where the end-to-end discovery workflow is iteratively optimized under the broad direction of scientists but without requiring step-by-step human oversight.[74] Once the agent system is trained, it can describe experiments necessary to test the generated hypotheses, analyze the results of said experiments, and use them to improve its internal scientific knowledge models. Agents in the self-driving system need to address the following three elements: determine inductive biases to reduce the search space of hypotheses, implement methods to rank order hypotheses considering their potential biomedical value with experimental cost, characterize skepticism via uncertainty quantification and analysis of experiments in reference to the original hypothesis, and refine hypotheses using data and counterexamples from experiments.[75] Ideally, hypothesis agents are creative and reflective when developing biological hypotheses that extrapolate indirectly from the existing body of knowledge.[16] There is emerging evidence that generative models have the potential to generate novel hypotheses. Tshitoyan et al.[76] demonstrated that using latent knowledge from published materials science literature can recommend novel materials. GPTChem[77] leveraged LLMs trained with an autoregressive pretraining objective to predict molecules. Experimental agents steer operational agents that use a combination of *in silico* approaches and physical platforms to execute experiments. Reasoning agents integrate the latest results to guide future experimental design. The utility of experimental results, such as the yield of high-throughput screening of a chemical library against a biological target, can be compared for different versions of the agent system given a time budget for hypothesis and experiment generation.

LEVELS OF AUTONOMY IN AI AGENTS

When integrated with experimental platforms, AI agents can operate at varying levels of autonomy tailored to diverse requirements across biomedicine. We classify these AI agents into four levels according to their proficiency in hypothesis generation, experimental design and execution, and reasoning (Table 1). Specific capabilities within each area define these levels, necessitating that agents exhibit the capabilities for a given level across all areas (an agent with level 3 capabilities in the experiment area but only level 2 capabilities in Reasoning and Hypothesis areas would be classified as level 2).

Level 0, denoted as "no AI agent," uses ML models as tools coordinated by interactive and foundation learning models. At

Table 1. Levels of autonomy in AI agents

Autonomy levels	Hypothesis generation	Experimental design	Reasoning	Human-AI collaboration
Level 0: no AI agents	none	ML models perform predefined tasks, with no adaptive changes to the protocols	none	scientist defines the hypothesis and sometimes uses the output of ML models to help with hypothesis generation; scientist defines the task to test hypothesis; scientist completes tasks
Level 1: AI agents as assistants	AI agent formulates simple and narrow hypotheses that are a direct composition of existing knowledge, preliminary data, or observations	narrow design of experimental protocols and utilization of *in silico* and experimental tools	strong reasoning in a selected task; multimodal summary of findings; use of experimental data and existing knowledge	scientist defines the hypothesis; scientist defines the series of tasks to test hypothesis; AI agent completes tasks
Level 2: AI agents as collaborators	AI agent generates hypotheses that are an explicit continuation of data trends and known literature	design of rigorous experimental protocols and adept utilization of a broad range of *ex silico* tools; once data are collected, employ statistical and computational methods to analyze the results and interpret the data to determine whether it supports or refutes the hypothesis	interpreting findings within existing knowledge, considering alternative explanations, and assessing the reliability and validity of the findings; synthesis of concepts beyond a summary of findings; collaborating with other researchers and undergoing peer review to validate findings and ensure that conclusions are robust and credible	scientist proposes initial hypothesis and refines hypothesis together with AI agent; AI agent defines the series of tasks to test hypothesis; AI agent completes tasks
Level 3: AI agents as scientists	AI agent generates creative, *de novo* hypotheses that are indirect extrapolations from existing knowledge	development of experimental methods unlocking new capabilities; actively gathering data through experiments or simulations using various techniques and tools to measure and record biological phenomena	based on the results and interpretations, refine experimental approaches for continuous learning and adaptation to improve the accuracy and depth of understanding; find concise, informative and clear conceptual links between findings	scientist and AI agent together form hypothesis; AI agent defines the series of tasks to test hypothesis; AI agent completes tasks

AI agents are characterized by four levels of autonomy in biological research, which are defined based on the capabilities of AI agents to complete different steps of the discovery process. At level 0, there is no AI agent, and ML is used as a tool. Level 1 consists of AI agents as research assistants, where agents complete a set of narrow and specific tasks defined by scientists. At level 2, AI agents act as collaborators and can use a broad set of tools to identify scientific discoveries. Still, they can only generate hypotheses that are a linear continuation of literature. Finally, at level 3, AI agents act similarly to human scientists across several axes of human evaluation, capable of identifying and understanding pioneering discoveries and extrapolating novel hypotheses that cannot be derived from existing knowledge.

this level, ML models do not independently formulate testable and falsifiable statements[78] as hypotheses. Instead, model outputs help scientists to form precise hypotheses. For example, a study employed AlphaFold-Multimer to predict interactions of "DONSON," a protein with limited understanding, leading to a hypothesis about its functions.[79] Level 1, termed "AI agent as a research assistant," features scientists setting hypotheses, specifying necessary tasks to achieve objectives, and assigning specific functions to agents. These agents work with a restricted range of tools and multimodal data to execute these tasks. For instance, ChemCrow[2] combines chain-of-thought (CoT) reasoning[80] with ML tools to support tasks in organic chemistry, identifying and summarizing literature to inform experiments. In another example, AutoBa[81] automates multi-omic analyses. These two agents are designed for narrow scientific domains; ChemCrow and AutoBa optimize and execute actions to complete tasks that are designed and predefined by scientists. Level 1 agents[2,81–83] formulate simple hypotheses inferred from existing knowledge and utilize a limited set of tools, lacking the capacity necessary to achieve level 2 autonomy.

At level 2, "AI agent as a collaborator," the role of AI expands as scientists and agents collaboratively refine hypotheses.

Agents undertake tasks critical for hypothesis testing, using a wider array of ML and experimental tools for scientific discovery.[76] However, their capability to understand scientific phenomena and generate innovative hypotheses remains constrained, highlighting a linear progression from existing studies. The transition to level 3, or "AI agent as a scientist," marks a major evolution, with agents capable of developing and extrapolating hypotheses beyond the scope of prior research, synthesizing concepts beyond summarizing findings and establishing concise, informative, and clear conceptual links between findings that cannot be inferred from literature alone, eventually yielding a new scientific understanding. While multiple level 1 agents exist across various scientific fields, levels 2 and 3 agents have yet to be realized. Existing taxonomies of autonomy focus on the division of responsibilities between AI agents and humans, with no consideration of biomedical discovery. These taxonomies were developed with the goal of advancing artificial general intelligence to surpass human performance across different skill levels, rather than being tailored to scientific research.[84]

As the level of autonomy increases, so does the potential for misuse and the risk of scientists becoming overly reliant on AI agents. While agents have the potential to enhance scientific integrity, there are concerns regarding their use in identifying hazardous substances or controlled substances.[85] Responsible development of agents requires developing preventive measures.[86,87] The responsible deployment of agents must account for the risk of over-reliance, particularly in light of evidence that LLMs can produce convincing but misleading claims and spread misinformation. The risks will likely increase as agents undertake more autonomous research activities. Agents must be scrutinized as scientists, including reproducibility and rigorous peer review of agentic research. We illustrate these definitions of levels by giving examples in genetics, cell biology, and chemical biology (Table 2). We selected these areas because of the availability of large datasets that have recently driven the development and application of ML models. Key ML and biological terms are described in Tables 3 and 4.

Illustration of AI agents in genetics

Research in human genetics seeks to understand the impact of DNA sequence variation on human traits. LLM-based agents operating at level 1 would perform specific tasks relevant to genetic studies. For instance, in a genome-wide association study (GWAS), a level 1 agent can write bioinformatics code to process genotype data to (1) execute quality control measures, such as the removal of single-nucleotide polymorphisms (SNPs) missing in many individuals or control for population stratification,[88] (2) estimate ungenotyped SNPs through imputation, and (3) conduct the appropriate statistical analyses to identify relevant SNPs, taking into account the false discovery rate.[89] Following the analysis, the level 1 agent reviews and reports findings, including any filtered SNPs and rationales for their exclusion.

Instead of executing narrow tasks following human instruction, a level 2 agent identifies and executes tasks independently to refine a hypothesis initially given by the scientist. For example, it may explore the effectiveness of drugs for a patient subgroup within complex diseases, where genetic underpinnings can influence drug response.[90] Given a hypothesis that a particular drug is effective in a subset of patients with idiopathic or genetic generalized epilepsy (GGE)—a condition with a robust genetic causality[91]—a level 2 agent would synthesize genetic information from GWAS meta-analyses,[92] such as the UK Biobank,[93] targeted sequencing studies,[94] and knowledge bases like Genes4Epilepsy.[95] The agent identifies GGE subtypes and causal genes by analyzing patient genetic data, predicting which subgroups might benefit from the drug based on genetic markers. It would then conduct *in vitro* functional studies to confirm these predictions, ultimately presenting evidence on how the drug could benefit GGE patient subpopulations by synthesizing concepts beyond summarizing findings.

Level 3 agents coordinate a system of agents (Figure 3) to discover and evaluate gene markers for specific phenotypes. These agents help initiate new study groups and optimize non-invasive methods of DNA collection for cost-effectiveness and recruitment processes.[96] Once data are collected, the agents innovate statistical methods to identify causal variants from genotypic data amidst confounders such as linkage disequilibrium[97] and develop *in vitro* techniques for validating candidate gene markers in disease models. Level 3 agents collaborate with scientists to generate and test hypotheses for comprehensive genetic insights.

Illustration of AI agents in cell biology

Cells are fundamental units of study in cell biology. Advances in single-cell omics, super-resolution microscopy, and gene editing have generated datasets on normal and perturbed cells, covering areas such as multi-omics,[98–100] cell viability,[101] morphology,[102] cryoelectron microscopy and tomography,[103,104] and multiplexed spatial proteomics.[105,106] This proliferation of data has spurred interest in *in silico* cell modeling.[107]

ML tools have been instrumental in analyzing data across these cellular modalities, but as level 0 agents, they lack autonomous research capabilities. At level 1, agents integrate specialized level 0 models to assist in hypothesis testing. These agents actively assist scientists in developing hypotheses by synthesizing literature and predicting cellular responses using integrated models. For example, to help investigate the resistance mechanism of a compound, level 1 agents predict its effects in various cellular contexts.[108] These predictions also inform experimental design, such as spatial transcriptomic[109] and proteomic[110,111] screening. Agents then retrieve and refine experimental protocols for execution on platforms[112] and apply predefined bioinformatics pipelines, as instructed by scientists.

Level 2 agents execute predefined tasks and generate hypotheses on cellular functions and responses. They autonomously define and refine tasks to support scientific reasoning, enabling practical exploration of complex phenotypes like drug resistance. By managing the experimental cycle and continuously updating their *in silico* tools, level 2 agents actively optimize experiments to focus on key variables of resistance based on a synthesis of predictive content, uncertainty, and newly acquired data, with iterative feedback from scientists.[55] Level 2 agents thus form a prototype for a virtual cell model capable of

Table 2. Examples of levels of autonomy of AI agents in genetics, cell biology, and chemical biology

Autonomy levels	Genetics (mutational effect modeling)	Cell biology (drug resistance)	Chemical biology (binder design)
Level 0	statistical package to analyze a pre-selected GWAS study	use of ML tools for modeling cellular outcomes of drug perturbations, including cell imaging, omics, and viability	use of ML tools for protein structure prediction, molecular docking, and generative models for binder design
Level 1	to explore potential mutational associations with disease, writes bioinformatics software for quality control and statistical analysis of genotype data from pre-fetched relevant GWAS studies	integrates multimodal (imaging, omics, and viability) and multiscale (cellular, tissue) data to create *in silico* models of drug resistance; retrieves and executes existing experimental protocols to study resistance; analyzes raw image and omics data with predefined pipelines	studies a specific protein target, integrates ML tools, such as AlphaFold for structure prediction and neural networks for screening chemical libraries to find candidate chemical compounds to bind to the target
Level 2	selects GWAS studies relevant to a provided hypothesis; if none exists, it designs and executes its own study or pulls other relevant genomic data to investigate the hypothesis	autonomously develops and adaptively refines hypotheses about resistance mechanisms based on knowledge and real-time experimental data analytics; designs and executes scalable and cost-effective experimental protocols with experts in the loop	designs binders for more challenging targets; identifies scaffolds that bind to similar pockets and adapts them for the target; synthesizes and tests molecules using existing experimental techniques
Level 3	initiates genomic studies and optimizes non-invasive methods of DNA collection for cost-effectiveness and ease of participant requirement; innovates statistical methods to identify causal variants from genotypic data and develops *in vitro* techniques for validating candidate gene markers in disease models	proactively identifies critical unresolved problems in drug resistance, proposing innovative therapeutic strategies; performs *in silico* simulations of cellular dynamics in tumor contexts and under complex perturbations (combinatorial genetic and chemical perturbations under different dosing schedules); develops novel highly multiplexed *in vivo* single-cell spatial technologies, enabling live tracking of gene expression, molecular interactions, and cell-cell interactions during resistance evolution	proposes *de novo* binders for an undruggable target or a poorly studied target; designs *in situ* experiments to study molecular interactions; synthesizes molecules with more complex pathways and designs and executes assays to test efficacy

Table 3. Glossary of key machine learning terms

Term	Description
Multimodal foundation model	advanced algorithms trained on multimodal datasets that can process various data types, including text, images, biological sequences, and high-dimensional tabular readouts; this training allows them to perform a broad array of tasks through few-shot fine-tuning and prompting across domains with little to no additional training
Transformer architecture	deep learning model architecture that uses on self-attention mechanism to capture long-range dependencies in input sequence data
Large language model	machine learning model with parameters on the scale of billions, trained on vast amounts of text data to understand, generate, and interact with human language on a large scale
Generative pretraining	strategy for training a machine learning model in an autoregressive manner to predict the next token from given data tokens, facilitating a general understanding of data sequence likelihoods
LLM-based AI agent	AI system capable of solving complex tasks within its environment by equipping large language model with modules for perception, interaction, memory, and reasoning
Embodied AI agent	AI agent system that interacts with the physical world through a body; the embodiment enables the agent to learn and adapt from sensory feedback and physical interactions
Fine-tuning	a training process of making small adjustments to a pre-trained machine learning model to improve its accuracy on a specific task or dataset
Instruction tuning	a training strategy that fine-tunes a model using a dataset of instructions and corresponding outputs to enhance its ability to follow specific instructions
Reinforcement learning with human feedback	a reinforcement learning strategy where an action model learns to perform tasks by receiving feedback from a reward model that mimics human preferences, guiding it to align with desired human behaviors
Prompting	techniques that provide specific text or other modal input instructions to guide the model in responding toward a desired output direction
Cross-modal alignment	a training scheme to align the representation embeddings of models across various modalities
In-context learning	ability to perform new tasks based on a handful of examples provided within the contextual prompt, without requiring explicit model training
Retrieval-augmented generation	techniques that make generative models to produce contextually relevant text by retrieving pertinent information and using it to inform the generation process

hypothesis generation, encompassing closed-loop integration of digital and experimental platforms.

Level 3 agents respond to existing challenges and anticipate future directions in cell biology research. They form hybrid virtual cell models by combining AI tools (digital agents) with high-throughput platforms (experimental agents). Digital agents, such as LLM-based agents, autonomously identify critical knowledge gaps through literature synthesis based on criteria such as data volume, biological relevance, and clinical needs and simulate any perturbagen (extrinsic events such as gene knockouts and overexpression, compounds, cell-cell interactions; intrinsic events such as cell cycle) in any context. Experimental agents not only optimize experimental protocols[102,113,114] to enable high-throughput multimodal measurements but also develop transformative technologies to enable probing at unprecedented resolution across space and time across *in vitro*, *ex vivo*, and *in vivo* models, uncovering pioneering insights. The ability of level 3 agents to drive the discovery of biological mechanisms and therapeutic strategies shifts the role of scientists from conducting operational tasks to focusing on ideation and managing hybrid cell models.

Illustration of AI agents in chemical biology

A major focus for chemical biology is understanding molecular interactions within cells to manipulate biological systems at molecular and cellular levels. An AI agent could analyze any molecular interaction, help design new drugs, and provide more valuable chemical probes for biological systems.

Despite considerable advances in applying ML to chemical biology, current approaches fall in level 0. Scientists oversee all activities by integrating ML tools for structure prediction, docking, chemical synthesis, and molecular generation. At level 1, the agent has elementary reasoning of chemical biology and can execute simple tasks autonomously, such as running ML tools or designing experiments for a given objective. However, due to limited reasoning capabilities, the agent may fail to explain more complex concepts, such as how the dynamics of molecules may influence the effects of drugs on binders or explore novel molecular scaffolds. For a level 2, the long-term objective is its function as a collaborator for scientists through excelling at tasks that are explicit continuations of existing scientific research, such as improving the efficiency of chemical probes, autonomously designing and testing *de novo* enzymes, or designing new binders by leveraging trends in related targets.

Table 4. Glossary of key biological terms

Term	Description
Linkage disequilibrium	a phenomenon in which two alleles occur so often in proximity in the chromosome that their association cannot be random
Single-nucleotide polymorphisms	genetic variation consisting of the replacement of a single nucleotide in the DNA sequence
Genome-wide association study	approach that identifies genetic variations across the entire genome associated with a specific disease or complex trait
Pharmacogenetics	field of research that aims to understand individuals' responses to different drugs based on their genetic factors
Experiment *in vitro*	procedures and investigations that occur within a laboratory environment (e.g., in a test tube) and outside of living organisms
In silico modeling	the use of computers to build simulations or experiments that recreate complex biological phenomena to be able to study and predict specific behaviors; for example, modeling of molecular dynamics
Mass spectrometry	analytical tools to characterize and identify individual molecules based on specific properties (e.g., mass-to-charge ratio)
Molecular docking	computational simulation tools used to predict how ligands bind to receptors
Retro-synthesis	techniques to design the synthesis of complex molecules by starting from the target and moving back to the original compounds
Crystallography	field of science studying the structure of atoms and molecules in crystals, which are solid materials whose compounds are ordered according to a very regular and ordered arrangement
Cryoelectron microscopy	imaging techniques used to identify the 3D structure of bio-molecules with near-atomic resolution without the need for extensive sample preparation and with the overall preservation of the sample
Single-cell RNA sequencing	high-throughput method that isolates individual cells and sequences their mRNA to measure gene expression levels of individual cells
Single-cell ATAC sequencing	high-throughput method that isolates individual cells and sequences their accessible chromatin to measure chromatin accessibility levels of individual cells

Level 2 AI agents have deeper expertise in more domains, such as retro-synthesis, crystallography, bioassays, and directing robotic arms to conduct research.

The goal of a level 3 agent in chemical biology is the ability to study all types of molecular interactions in a cell. This agent would work alongside human scientists to explore research questions that are challenging for the field, such as binder design for undruggable targets,[115] significantly improving specificity and efficiency of *in vivo* bioorthogonal reactions, or developing new chemical probes that can access new spatial and temporal scales. Unlike the level 2 agent's use of well-established protocols, a level 3 agent aims to unlock experimental capabilities that are not currently accessible. For example, AI agents could be tasked to probe molecular dynamics at longer timescales than what is currently accessible. At this level, agents have a thorough understanding of existing literature and work alongside scientists to unlock new fields of chemical biology.

ROADMAP FOR BUILDING AI AGENTS

An AI agent is built as a compound system that consists of modules[3,57,58] each implementing a distinct functionality. Here, we describe these modules (Figure 4), focusing on perception, interaction, memory, and reasoning modules necessary for AI agents to interact with humans and engage with experimental environments. Interactions between the agent and its environment are characterized by two elements: the agent's perception of its surroundings and its subsequent engagement with them. Perception modules enable the agent to interpret and assimilate information from various data modalities. Then, learning and memory allow agents to interact with an environment and complete tasks by acquiring new knowledge and retrieving previously learned ones. Finally, the reasoning module processes information and executes action plans. Using a published study as an example,[116] Figure 5E illustrates a hypothetical AI agent that sets up experiments to study the selective removal of mitochondrial DNA in *Drosophila* through perception, interaction, memory, and reasoning modules.

The division of research into smaller tasks handled by AI agents presents an intriguing approach, building on the success of modular and sequential bioinformatics workflows like Snakemake and Docker. Unlike these workflows, which are often static and require manual updates and reconfiguration to handle new tasks or integrate new tools, AI agents are dynamic and operate in a personalized, user-specific, and context-appropriate manner. They can learn to use new tools and adjust their workflows based on the specific instructions and needs of the scientist. Further, the adaptive allocation of tasks by AI agents can be helpful in automatically incorporating new tools and restructuring existing pipelines, much like a human researcher would. For example, AI agents could experiment with and create new protocols beyond the currently established methods in integrating multimodal omics data. For instance, while established protocols for integrating multimodal, such as single-cell RNA sequencing (scRNA-seq) with scATAC-seq or spatial data, exist, AI agents could develop new pipelines for multimodal integrations beyond the three modalities, or multiscale integrations such as atlas-scale single-cell and bulk RNA-seq data, or normal and disease state data from cell lines, organoids, and patient samples, based on their initial attempts.

Perception modules

Perception modules equip LLM-based agents with the capability to understand and interact with elements in the environment in which they operate, such as biological workflows and human users. For perception, agents need to integrate abilities to

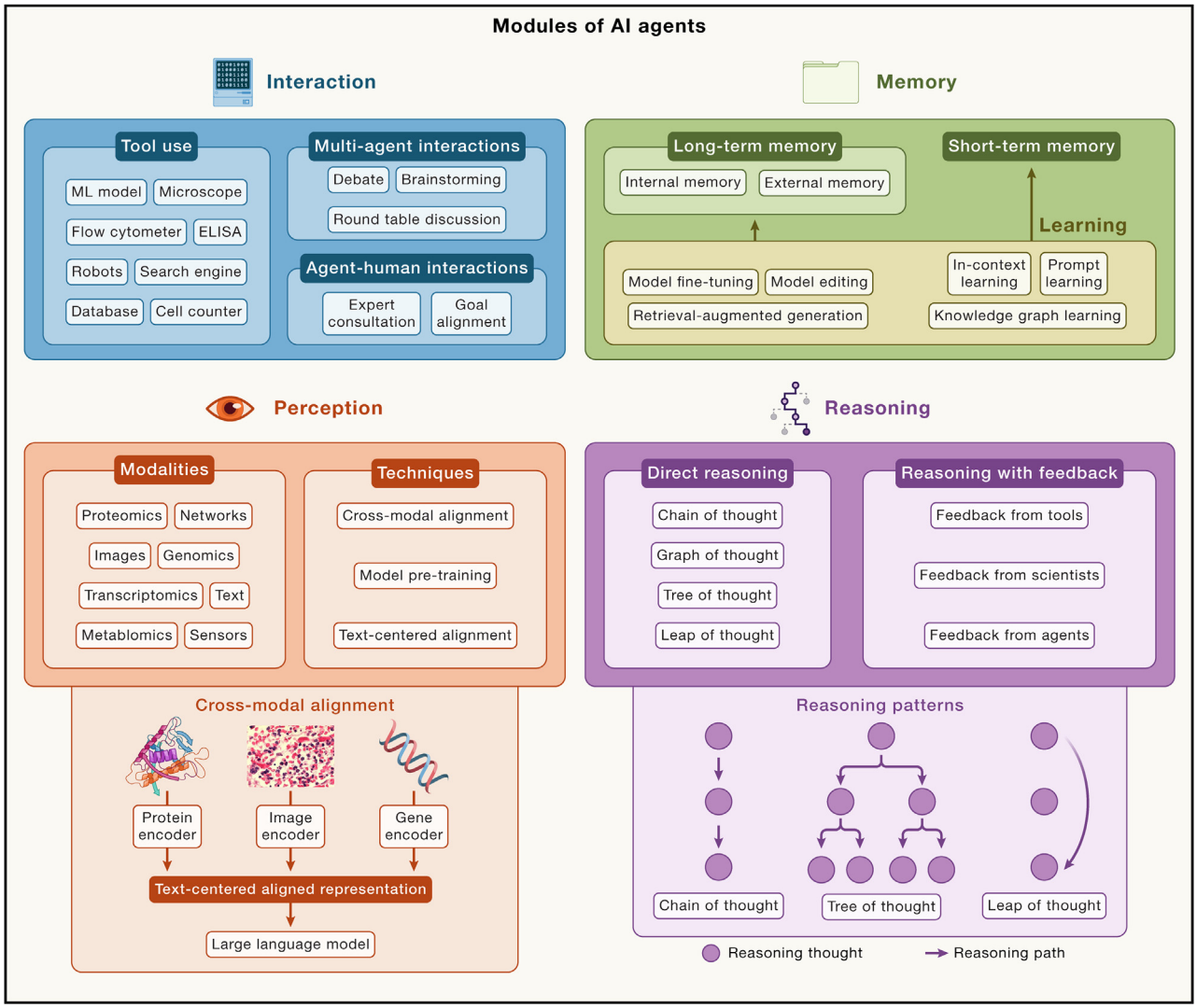


Figure 4. Key modules in AI agents: perception, interaction, reasoning, and memory modules

Perception interprets multimodal environmental data. Interaction facilitates engagement with the environment, encompassing human-agent interactions, multi-agent interactions, and tool use. Memory is responsible for the storage and retrieval of knowledge, while learning focuses on the acquisition and updating of knowledge. Reasoning, with or without environmental feedback, plays a crucial role in planning and decision-making processes. Cross-modal alignment is a key technique for the perception of LLM-based agents, where inputs from different modalities are aligned within a text-centered representation space. This alignment enables the LLM to perceive and process various input modalities. Reasoning patterns for AI agents indicate transitions between reasoning thoughts. For instance, agents with a chain-of-thought pattern generate reasoning in a step-by-step manner.

receive feedback from multiple sources: scientists,[49] the environment,[62] and other AI agents.[13,117] This requires accommodating a diverse array of modalities. These include text descriptions[6]; images from light and cryoelectron microscopy to assess cellular processes across many conditions simultaneously[103,104,118]; videos from live imaging to assess developmental processes or animal behaviors across time[119]; longitudinal biosensor readouts and genomics profiles of cells[120]; mass spectrometry-based proteomics to decipher protein homeostasis[24,121]; and miniaturized platforms for conducting biochemical assays and 3D culture systems that mimic the physiological context of organ systems.[112]

AI agents can take different approaches to interacting with environments. The most direct one involves using natural language, which represents a common perception modality for LLM-based agents. Other techniques involve multimodal perception modules, where agents process multimodal data streams from the environment or align multimodal inputs with text-based LLMs.

Conversational modules

With the rise of ChatGPT, the ability of AI agents to interpret natural language has reached such a high level[49] that it is now possible to build interfaces to agent systems that are entirely based on natural language with limited misinterpretations. The

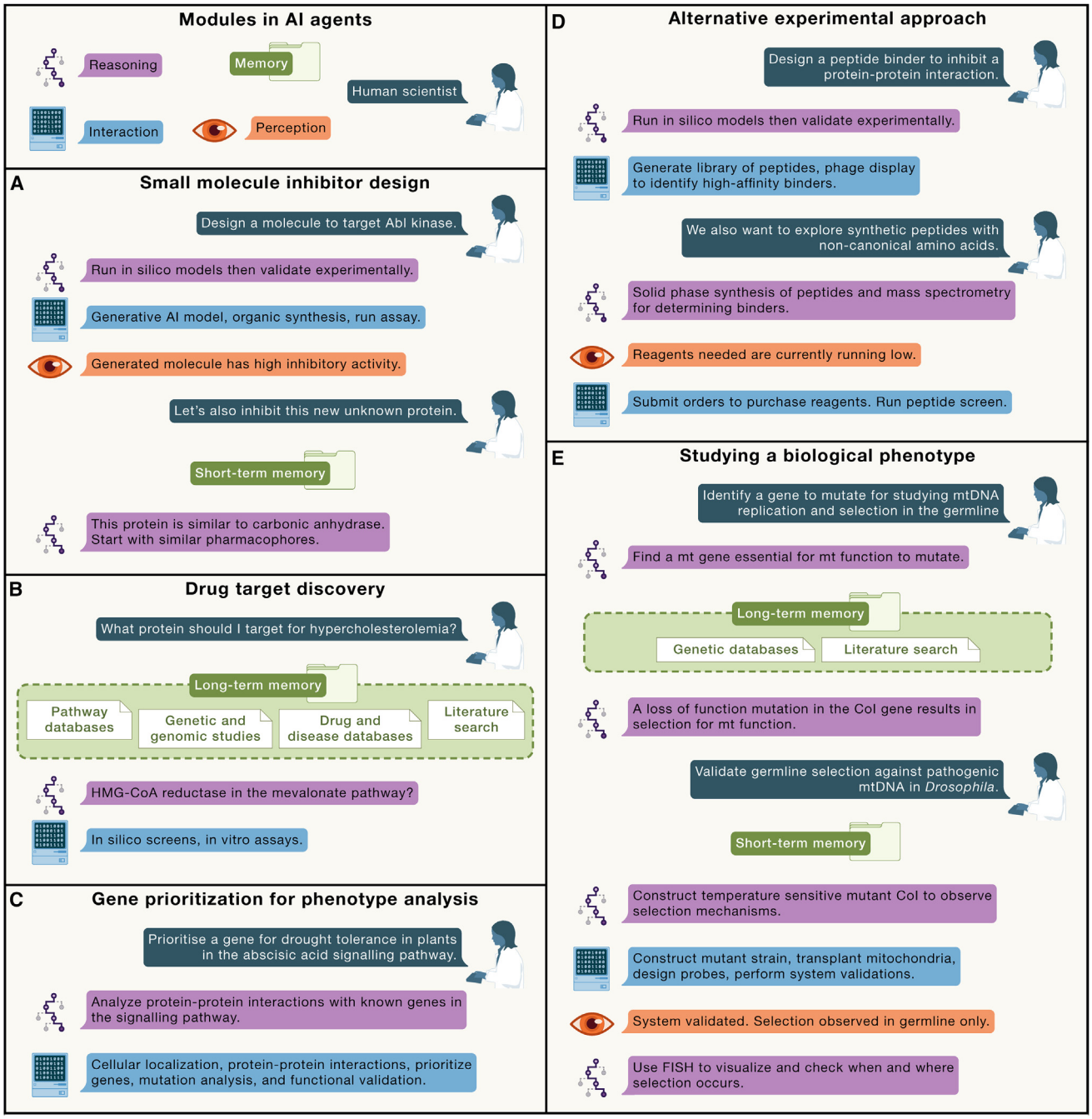


Figure 5. Illustration of components in biomedical AI agents

(A) Use of a short-term memory module to recall previous relevant experiments for small molecule inhibitor design.
(B) Use of a long-term memory module to retrieve relevant information for target prioritization for a disease.
(C) Use of direct reasoning without scientist feedback to prioritize genes for downstream phenotype analyses.
(D) Use of reasoning with feedback from scientists to select and optimize an alternative experimental approach.
(E) Combining perception, interaction, memory, and reasoning modules to study the selection against pathogenic mitochondrial DNA in the germline.

main focus is chat interfaces that preserve conversational history in a scrolling window, where users can converse with agents in a manner that resembles the standard approach of written human-to-human interaction. This approach allows scientists to express their queries using their language, promoting initiative and enabling them to precisely describe what they want. We envision that agents will maintain a history of interaction with scientists using natural language, which, in turn, will allow us to keep track

of scientific interactions with agents.[62,68] Combining traces of these interactions with RAG, we can develop personalized discovery workflows tailored to individual scientists.

Multimodal perception modules

Agents align LLMs with other data types to fuse data modalities beyond natural language text. This approach helps agents better model the changing environment in which the agent acts and dynamically adjust its outputs to new situations, such as evolved biological states in a virtual cell model. The alignment process involves two main strategies: textual translation and representation alignment. Textual translation converts inputs into a textual format, such as transforming data from robotics into textual descriptions that log environmental states.[9] For example, when handling readouts from experimental devices, the readouts can be combined with a textual description of their meaning, allowing the LLM to understand the readouts as a new modality. Alternatively, through representation alignment, data from different modalities are analyzed by modality-specific models to generate representations, such as using the visual encoder from CLIP[122] for visual information processing. These representations are then aligned with LLM textual representations through instruction tuning,[118,123] enabling agents powered by LLMs to perceive and interpret multimodal data. For instance, to make LLM-based agents handle the protein structure data, an additional encoder is required to encode the protein structure data into a representation aligned with the LLMs' representation space. This encoder is pre-trained with modality-specific training schemes, and an adaptor is placed between this encoder and LLMs to align the representations of the two modalities. Then, instruction tuning is applied using data containing both modalities to train the adaptor for alignment. An alternative to alignment involves allowing the agents to receive input expressed in different modalities.[7,124] For instance, Fuyu[124] uses a decoder-only transformer architecture to process image patches and text tokens jointly. Similarly, Gemini[7] is engineered to handle visual, audio, and text inputs within a single model. Once perception modules are implemented for agents to receive inputs from the environment, modules for interaction and reasoning follow to process the inputs and interact externally. Training agents with strong perception abilities on biomedical data requires extensive, high-quality data pairs that align multiple modalities. However, collecting such data remains challenging. For example, multimodal experimental platforms are non-existent or have low-throughput yields, certain tissues and cell types are not experimentally available, and a long tail of disease phenotypes has small sample sizes, making data collection infeasible.

Interaction modules

Beyond conversational modules, scientists use ML-based and other tools in biological research, explore datasets through graphical user interfaces (GUIs) to analyze and visualize data, and engage with physical equipment and wet lab experimental platforms. Chat-optimized LLM-based agents thus need interaction capabilities to communicate and collaborate with scientists, other AI agents, and tools to function beyond a simple chatbot. Agents must incorporate essential interaction modules to interact with elements in the environment. These include agent-human interaction to support communication with scientists and following human instruction,[125,126] multi-agent interaction for collaboration among agents, and tool-use action to access ML tools and experimental platforms.

Interactive abilities of LLMs, when combined with function calling, can act as an intermediary between scientists and the agent's interface, as well as between scientists and other functional items, such as tools and other agents. This approach allows scientists to express their intentions in natural language without needing to search for how and where to accomplish tasks. At the same time, the advantages of functional items are preserved because agents can interact with tools and use them to provide feedback. However, interactive modules trained on general, non-biological domains might not be well-suited for specialized biomedical terminologies, requiring in-domain training on biomedical tools.

Agent-human interaction modules

The interaction between scientists and AI agents synchronizes scientific objectives with AI agents through cooperative communication and modeling of biological knowledge. Natural language processing and human evaluation methods are predominantly used to develop this interaction capability. InstructGPT[49] enhances the GPT model through supervised fine-tuning with examples of human dialogues to improve the model's conversational skills. The alignment between agents and humans can be refined through RLHF, which adjusts the model based on a reward model trained using human assessments of the model's responses. Alternatively, RLHF can be replaced by direct preference optimization,[127] which is a parameterized method that provides a more consistent and efficient alignment with human preferences. Through agent-human interaction, agents become attuned to human needs and preferences,[10,126] using human insight as a directive for carrying out complex tasks.[13] For instance, Inner Monologue[126] employs human feedback to discern user preferences or interpret ambiguous requests in an embodied context. In AutoGPT,[10] humans formulate tasks and score solutions returned by agents, and AutoGen[13] can use human expertise to solve tasks better than agents alone.

Multi-agent interaction

Multi-agent interactions support solving complex goals that agents could not complete if they operated independently. In such interdisciplinary systems, agents that could specialize in different biological domains, each with distinct capabilities, engage in interactions through various communication means. Language has emerged as the predominant medium for multi-agent interactions due to the ability of agents to communicate with humans linguistically.[4,13,73,117,128] An instance of this is generative agents,[62] which create interactive environments where agents mimic human behavior and interact using natural language. Different strategies are used for multi-agent interaction, including cooperation[129–131] and negotiation.[73,132,133] For example, MetaGPT[130] applies standardized operating procedures from human teamwork to define tasks and agent responsibilities.

Through these approaches, agent interactions make it possible to tackle tasks that are too complex for just one agent to handle.[82,134] MedAgent[82] leverages the expertise of multiple medical AI agents for medical reasoning. Similarly, RoCo[134] employs robot agents with varied roles to accomplish complex

tasks in the physical world. Multi-agent interaction can also boost the proficiency of less skilled agents by allowing them to learn from more experienced counterparts.[135] These interactions also enable the creation of simulations for a variety of environments, ranging from public health scenarios[136] to human social behaviors,[62,137] enhancing the system's adaptability and application in diverse contexts.

Tool use

To manage tasks from diverse environments, agents require tools to boost their capabilities.[138] Commonly used tools are application APIs,[139] search engines,[140] ML models,[141] knowledge DBs,[142] and robotic machinery for physical tasks.[9,143,144] Level 1 agent systems have been developed that can interact with one or more types of tools. ChemCrow[2] leverages chemical tools and search engines to address chemical challenges. WebGPT[140] can conduct searches and navigate web browsing environments. SayCan[144] controls a robot in the physical world using an LLM to complete tasks. To invoke these tools, AI agents generate commands in specific formats[139,141,142] or query pretrained control models to execute actions.[144,145] To develop these capabilities, agents can use in-context learning[141] or fine-tuning with tool-use demonstrations,[139] where the latter represents a more sophisticated approach.

For in-context learning, it is necessary to include system abilities in the prompt so the agent can use function calling to query tools. For example, HuggingGPT[141] uses ChatGPT as a controller to integrate all ML models on Hugging Face through in-context learning. The alternative approach consists of using model fine-tuning with function calling to create an LLM-based agent with integrated abilities of a function/tool. For instance, Toolformer[139] introduces a self-supervised learning method to master the use of tools' APIs with minimal demonstrations for each API.

By modeling scientists' needs by analyzing natural language textual inputs, AI agents can select the most likely available tool, identify the desired user interface component, and execute the scientist's expected actions. Interaction modules are designed to be integrated and adapted to suit changing environments. For level 2 and level 3 agents, agents autonomously learn new types of interactions and how/when to start using new tools.

Memory and learning modules

When using tools and ML models for biological research, scientists keep records of experimental logs and plan their next steps based on them. In AI agents, memory modules alleviate the need for manual log recording by memorizing necessary experimental outputs. Contrary to ML models that perform one-time inference to generate predictions, memory modules in LLM-based agents store and recall information. This is necessary for executing complex tasks and adapting to new or evolving environments. Memory modules are designed to store long-term and short-term learned knowledge. As agents encounter new situations and acquire data, memory modules get updated with new information.

Long-term memory modules

Long-term memory stores essential and factual knowledge that underpins agent behavior and understanding of the world, ensuring this information persists beyond task completion. This memory can be internal, encoded within the model's weights via learning processes,[8,146] or external, maintained in auxiliary knowledge bases.[147,148] Internal memory is directly used for accomplishing zero-shot tasks[6,7] while accessing external memory requires actions by the agent to fetch and integrate data into short-term memory for immediate use.[149,150] For instance, ChatDB[142] uses an external DB for memory storage, and MemoryBank[151] encodes memory segments into embeddings for later retrieval. Agents can query knowledge banks, such as a GWAS DB to find genetic evidence for a candidate protein target, a knowledge base of therapeutic mechanisms of action, and scientific literature with up-to-date information for the agent to integrate and decide whether the protein can be modulated through a therapeutic perturbation (Figure 5B). The learning process updates long-term memory by adding new knowledge or replacing outdated information. Internal memory of an agent can be updated using parameter-efficient fine-tuning,[146,152] interactive learning,[49] and model editing.[153] These strategies must be effective for large models[152] and avoid the loss of previously learned information.[154] On the other hand, updating external memory is more straightforward, involving modifications to the knowledge base.[142,151] For example, in drug discovery, updating long-term memory by adding a new compound in development to the drug bank is a convenient way to maintain an up-to-date agent.

Short-term memory modules

AI agents use short-term memory to temporarily store information during their interactions. This short-term memory is enabled through in-context learning, where relevant information is integrated as context prompts[144,155] or via latent embeddings[118,123] in LLMs. For chatbots, previous conversations are kept as text prompts, supporting multiple rounds of dialogue.[49,156] The text-based approach lays the groundwork for communication in multi-agent[73,133] and agent-human scenarios.[10,13] In embodied AI agents, environmental feedback[144,155] is captured in textual format, acting as a short-term memory that aids reasoning. Following perception, multimodal inputs are converted into latent embeddings, which function as short-term memory. LLaVA[118] uses latent embeddings generated by visual encoders to retain visual information. Short-term memory allows agents to temporarily acquire skills, such as tool usage,[139,141] store information about recent states of a biological system,[155,156] and keep track of outcomes from earlier reasoning efforts.[11] This learning mechanism is crucial for agents to learn and apply new knowledge under new conditions. Moreover, short-term memory can temporarily override long-term memory, allowing agents to precede recent information over older knowledge within their model weights.[157] Agents can be informed by past experiences stored in their short-term memory to tell which experiments to run in the future. In Figure 5A, we detail an example where the agent recalls experiments for a similar protein to inform the initial inhibitor design for the given protein.

Reasoning modules

Biological research involves a multidisciplinary and multistage process that integrates the expertise of scientists from various disciplines. Scientists formulate hypotheses, design experiments based on these hypotheses, interpret the results, and plan the

next steps. The integration of reasoning capabilities in AI agents can assist biological research throughout this process. Reasoning improves agents' capabilities to plan experiments, make decisions on biological hypotheses, and resolve competing candidate biological mechanisms. AI agents that use LLMs can implement interactive dialogue systems to explain ML models through natural language conversations. Reasoning modules can be implemented using prompting[158] and few-shot in-context learning.[80] Additionally, agents can use planner models[159,160] and action models.[155] We classify reasoning modules into two categories: direct reasoning and reasoning with feedback, depending on whether agents adjust their plan in response to experimental or human feedback.

Direct reasoning modules

In direct reasoning, an agent performs planning and reasoning based on the current state of the environment, which can follow different reasoning patterns, such as single-path and multi-path reasoning. Single-path reasoning involves the agent breaking down the task into multiple recursive steps.[161] For instance, CoT reasoning allows agents to reason step-by-step either by using in-context examples[80] or by applying a zeroshot prompt like "Let's think step-by-step"[158]. Leap-of-thought[162] encourages the model to use creative rather than logical reasoning. Although single-path reasoning matches well with certain situations,[163] its ability to adjust to different conditions is limited.

Conversely, multi-path reasoning examines several paths before consolidating them into a final plan,[164,165] allowing for a more thorough planning process that accounts for different scenarios. For example, least-to-most prompting[166] breaks down tasks into subproblems solved sequentially. Self-consistent CoT[167] chooses the most consistent answer from a set of CoT answers. Tree-of-thoughts[164] extends reasoning paths into a tree-like structure, generating multiple paths from each thought node and using search algorithms to select the final path. Graph-of-thoughts[168] further develops reasoning paths into a graph structure for complex reasoning. To identify the optimal path, methods such as voting strategies,[167] Monte Carlo tree search,[169] and breadth/depth-first search algorithms[164] are used. Through direct reasoning, agents can generate thought that could consider the protein targets in a pathway and experiments to test the role of a candidate protein target (Figure 5C).

Reasoning with feedback

Experimental and human feedback can help AI agents to improve reasoning and planning processes.[11,68,149] This feedback may include agent-human interaction and responses from agents, which can be complementary biological assays quantifying downstream effects of target molecules.[170] In each reasoning cycle, React[11] incorporates insights from previous actions to refine its thought process and inform future actions. LLM-Planner[171] dynamically adjusts plans based on new observations in an embodied environment. Inner Monologue[126] uses both passive and active scene descriptions and feedback from recent actions to guide future actions. Voyager[68] improves planning for subsequent steps by considering environment feedback, execution errors, and self-verification.

Beyond external feedback, an agent's feedback mechanism enables self-assessing the initial plan.[170,172] Techniques like self-refine[170] revise action outputs based on the LLM evaluation, the self-check[170] mechanism allows the agent to review and adjust its reasoning, and reflection[12] mechanisms use prompt agents to update their decision-making. These techniques incorporate feedback from biologists, such as exploring experimental methods and environmental constraints like lab inventory (Figure 5D). Reasoning capabilities are necessary for generating hypotheses and conducting experiments. Generating novel hypotheses requires modeling general biomedical knowledge, the specific information on the current state of a biological system, and consideration of potential next steps. LLM-based agents can generate hypotheses through in-context reasoning, but careful selection is necessary to ensure high-quality hypotheses.[173]

CHALLENGES

This perspective outlines key steps for implementing AI agents in biomedical research and identifies areas that could benefit from agentic AI. However, challenges remain and may be amplified with the introduction of multi-agent systems (Figure 6).

Robustness and reliability

A barrier facing the deployment of agent systems—specifically those categorized within levels 2 and 3 as discussed in Table 1—is their propensity for generating unreliable predictions, including the hallucination of non-factual information, reasoning errors, systematic biases, and failures in planning when connected with tools and experimental platforms. These issues can be exacerbated by overconfidence in such flawed predictions (agents lack awareness of their knowledge gaps) and high sensitivity to the precise formulation of queries, particularly in the context of LLM-based agents. This behavior has been traced to how these models are trained. In particular, autoregressive loss compares the predicted word sequence with the actual sequence in the training data. The performance of a model trained with this loss is determined by three factors: the probability distribution of the inputs, the sequence of generated outputs, and the frequency of different tasks encountered during training.[174] As a result, model performance degrades on task variants that deviate from the assumptions made during training.[175]

Sensitivity to input and task probability also offers a potential explanation for the widely observed success of various prompting techniques[80,164,176] (methods to paraphrase the same query). By providing informative context, instructive reasoning steps, or representative examples, these techniques can act as an empirical means by which task and input probability (and, thus, model performance) are increased. However, crafting high-quality prompts tends to be highly empirical while requiring significant effort and domain knowledge.

Beyond the linguistic domain, even the most advanced models fail in tasks with real-world entities that require physically meaningful actions, posing an obstacle to embodied agents. While embedding continuous sensor data into a language model can lead to improvements,[120] limitations to understanding physical interactions and long-horizon planning remain. The complexities of training such multimodal systems, the need for large datasets to cover the range of embodied tasks and environments, and the computational demands of processing

Challenges for AI agents	Strategies to address challenges			
	Algorithms and software	Experimental platforms and hardware	High-fidelity datasets	Science and policy makers
Robustness and reliability				
Hallucination prevention	X			X
Embodied reasoning	X	X		
Predictive uncertainty and model understanding	X		X	X
Evaluation protocols				
Holistic evaluation protocols	X	X	X	X
Inherent variability of biological systems		X	X	
Standardized biomedical discovery workflows	X	X	X	X
Dataset generation				
Multimodal, noisy, and incomplete biological data	X	X		
Datasets optimized for AI model training			X	X
Governance				
Regulation and reporting standards	X		X	X
International collaboration and scientific consensus		X		X
Risks and safeguards				
Lab safety protocols for autonomous systems		X		X
Safety monitoring and certification	X	X	X	X

Figure 6. Challenges for AI agents in biomedical discovery

Shown are critical challenges—including robustness and reliability, evaluation protocols, dataset generation, governance, and risks—alongside strategic approaches to address them.

multimodal inputs all remain open questions.[7] Deployment faces challenges from false negatives causing repeated attempts and eventual stalling of the embodied agent.[126] Hence, it is necessary to verify the agent action plan before execution.

Uncertainty quantification can trigger fall-back safety measures like early termination, predefined safe maneuvers, or human-in-the-loop interventions. However, foundation models cannot reason about the uncertainty associated with their outputs, and no well-established statistical protocol exists for increasingly ubiquitous architectures.[47,177] Techniques such as various forms of prompting, e.g., Wang et al.,[167] Tian et al.,[178] and Kuhn et al.[179] estimate uncertainty based on the model's predictive distribution, p(output|input), which may itself be subject to bias[174]; furthermore, it does not consider the distribution of model parameters consistent with the observed training data and marginalizes over its predictions.[180] While conformal prediction[181] has emerged as a framework for uncertainty estimation of model predictions, its sensitivity to the choice of underlying statistical assumptions and the calibration of confidence levels have been criticized. The lack of a default technique is partly due to the difficulty of establishing a thorough quality assessment of uncertainty estimates. This makes it difficult to make choices in agent design and to reassure users about its calibration.

One concern is that advanced capabilities come at the cost of compromised transparency and the risk of misalignment. For instance, integrating human feedback can promote desirable agent behavior, but it can also exacerbate persuasive abilities, echoing false beliefs.[182] Fine-tuning existing models with new data can compromise their original alignment, challenging the integrity of the AI agent's intended purpose.[183] Jailbreak attacks can similarly affect post-deployment, highlighting the need for rigorous evaluation.[184]

Errors are inevitable in complex multi-agent systems, making their management crucial to maintaining system robustness and reliability. Due to their interactive nature, these systems are sensitive to compounding errors, where small issues can escalate into significant problems if not addressed promptly. Effective error management strategies are essential for diagnosing, localizing, and mitigating such errors.

Evaluation protocols

With more AI agents being developed, frameworks for biologists and lay user evaluations need to assess axes of agent performance beyond accuracy. Evaluating AI agents requires an analysis of their theoretical capabilities and an assessment of practical implications, including ethical considerations, regulatory compliance, and the ability to integrate into discovery workflows. The challenge lies in developing evaluations that consider these diverse factors. Agents that integrate ML tools, particularly those developed by corporations, may undergo updates without prior notice to users. This poses challenges for reproducibility, as updates may alter the model's behavior or performance without researchers being aware. The scientific community needs transparent change logs and version control for agents, akin to practice in software development.

Existing evaluation frameworks consider either holistic evaluations[185,186] or benchmark the models for weak spots such as task framing,[187,188] long temporal dependencies, invalid formatting, or refusal to follow instructions.[189] A caveat of such frameworks is the risk of evaluating how well the agents have learned to use specific APIs versus general results grounded in real-world interaction. Another challenge in evaluating agents is that biological systems are inherently dynamic, characterized by non-stationary distributions that evolve due to genetic mutations, environmental changes, and evolutionary pressures. Agents trained on static datasets may struggle to accurately model or predict outcomes in these changing systems. The challenge lies in

developing agents capable of adapting to or continuously learning from new data, ensuring their predictions remain accurate as the underlying biological systems change. Techniques such as online learning, transfer learning, and reinforcement learning can be used to address this issue, but they come with their own set of challenges related to data availability and model complexity. Another challenge is the lack of standardization in biomedical discovery workflows, including data generation protocols that vary based on factors like disease cell lines, dosage levels, and time points.[190] This variability complicates the evaluation of agents for experimental planning. Evaluation of agents that use computational tools and DBs will benefit from the increasing availability of standardized APIs.[191,192]

Dataset generation

As laid out, the vision for biomedical AI agents requires the capability of seeking, aggregating, perceiving, and reasoning over data from various modalities, created using differing specifications and with inherent variation in quality and volume. To support this vision, there is a critical need for large, open datasets that are both comprehensive and accessible, enabling the development of models across biological applications. Much human effort in building systems for biomedical research is dedicated to gathering and preparing such data for use in ML models (e.g., specific to a particular modality, such as graphs, time series, or discrete sequences[193]). This requires vetting processes and clear criteria for assessing the reliability and applicability of datasets.

Noisy data, characterized by errors, inconsistencies, and outliers, poses a significant challenge for models attempting to extract meaningful patterns and insights with minimal human oversight or data preparation effort. In addition, multimodal data require models to process different data representations and formats and bridge semantic gaps between them. Tackling these challenges necessitates advanced feature extraction, fusion, and noise mitigation techniques while maintaining robustness. As no pretraining phase (no matter how extensive) will be able to provide adequate examples from all data sources, models will also have to generalize to previously unseen sensory inputs.

Governance of AI agents

The governance of AI agents presents challenges that intersect technological, scientific, ethical, and regulatory domains. One challenge is establishing comprehensive governance frameworks that balance innovation with accountability.[194] As AI agents gain autonomy, the necessity for robust guidelines to ensure responsible development, deployment, and commercialization grows. The discourse increasingly advocates for agent safeguarding to take precedence over further advancements in autonomy. Yet, navigating the regulatory landscape and forging an international consensus on AI governance remains complex while the advancement of agent capabilities continues. Striking a balance between innovation and safeguarding against potential risks requires collaboration among industry leaders, scientists, and policymakers.[195]

Safe adoption of AI agents requires addressing concerns of safe deployment. Aligning ML tools, such as LLMs, with ethical standards remains an open challenge, and ensuring the alignment of the agent as a digital entity raises complexity. Guidelines concerning human-agent interactions are underdeveloped despite the potential for unintended harmful consequences and malicious intent. Safeguarding frameworks are developed that include training, licensing, and mandatory safety and ethical compliance checks for agents.[86]

As AI agents become more integral to workflows in biological domains, monitoring their behavior grows increasingly complex. Currently, verifying the accuracy and trustworthiness of agent outputs is not straightforward, with only a limited number of systems capable of linking generated content to relevant references. It is essential to develop robust verification systems that can provide traceable references for generated content. Assessing the synthesized knowledge may be impractical and unattainable as agents evolve further. When agents' capabilities become comparable to those of human experts, the risk of becoming overly reliant on AI increases, which could lead to a decrease in human expertise. In the worst-case scenario, such reliance could introduce a broad spectrum of safety hazards due to inadequate oversight. To address these challenges, human-in-the-loop approaches can help maintain accountability. Continuous training and development of human expertise alongside AI can mitigate the risks of over-reliance on AI.

Risks and safeguards

Autonomous experiments that do not include careful planning, broad consultation, competent execution, and ongoing adaptation might create long-term harms that outweigh the benefits. Although anticipating all potential complications is impossible, exploring possible problems early and frequently could reduce the expected cost of such issues. The ethical and technical considerations relevant to AI agents are vast and deeply interconnected, particularly in biomedicine. This section will highlight some key categories.

Neglect can lead to risks similar to those of malicious intent. Multi-agent systems where some agents represent LLMs might, through equipment malfunctions and insufficient maintenance, inadvertently create harmful substances, for instance, by contaminating a procedure that would otherwise be safe. This issue is not unique to multi-agent systems; instead, it is a general lab safety concern. However, the absence of close human supervision removes a critical auditing layer. The increased role of automation in agent systems raises safety issues: a powerful, unaligned system prone to misinterpreting user requests or unfamiliar with lab safety practices could, given access to a well-stocked scientific facility, do damage by, for instance, mixing volatile substances or developing and dispersing toxins or pathogens. These are among the scenarios that most concern AI safety researchers.

Agents leverage LLMs' world knowledge and general reasoning abilities obtained during pretraining for robotics and planning. However, while efforts have been made to teach the robots the "dos," the "don'ts" received less attention. Teaching robot agents the don'ts is crucial for conveying instructions about prohibited actions, assessing the agent's understanding of these restrictions, and ensuring compliance.[196] For LLM

agents, plug-in safety chips[196] feature safety constraint modules that translate natural language constraints into formal safety constraints for the robot to adhere to. Experiments with robots highlight the potential for integrating formal methods with LLMs for better robotic control.

LLMs trained in code completion can write Python programs from docstrings[197] by training the model on the code completion task to write the code based on natural language commands.[198] Given natural language commands, these code-writing LLMs can be repurposed to write robot policy code. However, if the translation inaccurately reflects the intended safety constraints, it could lead to either overly restrictive behavior, preventing the robot from performing its tasks effectively, or insufficiently stringent constraints, leading to safety violations. However, the robot policy code is less reliable for enforcing safety constraints than verifiable safe operations that satisfy standards such as International Organization for Standardization (ISO) 61508. The approach assumes that all given instructions are feasible and lacks a mechanism to predict the correctness of a response before execution. However, due to their reliance on patterns in the training data, LLMs might generate syntactically correct but semantically inappropriate code. Additionally, generalizing plans across robotic embodiments is brittle with current LLMs.

Addressing the ethical implications of AI agents is paramount, given the direct impact on human and animal health and life. The handling of sensitive biological and medical data necessitates robust technological and regulatory measures to ensure security and confidentiality. One promising approach involves using privacy-preserving computation to train agents to protect the privacy of highly sensitive medical data. Homomorphic encryption can secure sensitive data by allowing computations on encrypted data, and federated learning techniques allow training agents in a distributed manner without the need to centralize from across sites into a single data repository.

Algorithmic fairness is equally crucial, as biased AI agents can exacerbate health disparities across patients and increase inequalities in the volume of generated datasets and quality of biomedical knowledge, especially for diseases in long-tailed distributions in biological systems. The development of techniques such as adversarial debiasing and fair representation learning offers promising avenues to mitigate these risks. In addition, the black-box nature of these compound AI systems poses another challenge, particularly in healthcare, where interpretability is vital for clinical adoption and patient trust. To provide clearer rationales for the agents' decisions and make them more acceptable to users, it will become crucial to incorporate interactive dialogue systems that explain agentic outputs through natural language conversations. Ethical considerations surrounding biosafety emerge as AI agents advance toward level 3 agents. These issues intersect with ongoing debates in bioethics regarding synthetic biology and artificial organisms, requiring regulatory guidance and engagement from bioethicists and safety experts to ensure alignment with societal values and safety standards.

Challenges uniquely relevant for biomedical AI agents

Biomedical AI agents face several unique challenges that distinguish them from other applications of AI. While strong AI agents have the potential to mitigate some of these challenges, their implementation in biomedical research requires careful consideration. One of the primary challenges is the need for robust and reliable systems capable of reasoning, planning, and executing actions in both virtual and hybrid virtual-physical environments. For instance, natural language reasoning chains can enhance the interpretability of an agent's actions and contextual outcomes, aiding researchers in understanding AI-generated insights. However, certain challenges persist that can delay the reliable implementation of AI agents or even cause harm if these systems are deployed prematurely. A critical issue is the difficulty in distinguishing between correlation and causality. Current AI agents struggle with generating strong hypotheses, reasoning, and conducting experimental validations, tasks that typically require advanced AI systems (level 3 agents) or human intervention. Moreover, AI agents need improved interfaces to interact safely and effectively with experimental platforms. These platforms themselves face limitations in producing unbiased, AI-ready datasets that accurately capture the intra- and inter-variation inherent in biological systems. Such limitations hinder the generalization capabilities of AI agents, which rely on comprehensive and high-quality data to function optimally. The absence of data from high-throughput techniques can lead to AI agents forming false hypotheses or causing harm. This risk is exacerbated when AI agents work with small, biased biological datasets, which may be affected by issues like batch effects.

OUTLOOK

Biomedical research is undergoing a transformative era with advances in computational intelligence. Presently, AI's role is constrained to assistive tools in low-stake and narrow tasks where scientists can review the results. We outline agent-based AI to pave the way for systems capable of reflective learning and reasoning that consist of LLM-based systems and other ML tools, experimental platforms, humans, or even combinations of them. The continual nature of human-AI interaction and building trustworthy sandboxes,[199] where AI agents can fail and learn from their mistakes, is one way to achieve this. This involves developing AI agents proficient in various tasks, such as planning discovery workflows with ML feedback loops for experiments and performing self-assessment to identify and seek out gaps in their knowledge.

Ensuring context-appropriate and user-specific agent behavior

To ensure agents behave as intended, it is essential to focus on their robustness and reliability by implementing evaluation protocols that test agents in diverse scenarios to identify potential vulnerabilities. Moreover, grounding agents in ethical guidelines and documentation, such as lab protocols and safety guidelines, is vital to align their actions with human values and safety standards. By addressing these aspects, we can ensure that the behavior of biomedical agents is both reliable and ethically compliant.

Concretely, we believe that in the early stages of technological adaptation, it is desirable to limit an agent's capabilities to a subset of their full potential by restricting action spaces, thereby

eliminating the chance of catastrophic risk (e.g., decisions resulting in loss of life). Similar precedents for technological adaptation are already in place for other autonomous systems with similar risk profiles, such as autonomous driving, where a staggered technological adaptation is motivated by ethical considerations.

Governance and responsible human-AI partnership

Managing errors requires designing strategies to diagnose, localize, and mitigate them. To diagnose errors internally, agents should use their reasoning abilities to build self-evaluation schemes, allowing them to assess their current status and actions. Externally, training independent anomaly detection and distribution shift models with domain knowledge of specific biomedical use cases can provide additional supervision to diagnose errors. Iterative agent interactions can result in cascading errors. To mitigate this, the evaluation agent can apply reverse reasoning chains to trace back to the initial error. Enhancing the adaptive reasoning abilities of agents is crucial for dynamically adjusting to changing conditions and rectifying errors as they occur.

To address the challenge of governance, we believe that broad consensus is best achieved through multidisciplinary, cross-partisan, non-profit, and public institutions committed to the public good. We welcome the recent establishment of several public AI safety institutions to facilitate these discussions. Focus groups with expertise in AI agents can develop ethical and technical evaluation standards that can form the basis for regulation, including the required degree of human oversight and accountability frameworks. Additionally, we advocate for the development of policies through international initiatives to minimize the risk of regulatory gaps, where risks might otherwise be outsourced to jurisdictions lacking enforceable regulations.

By fostering responsible human-AI partnerships and establishing robust governance frameworks, we can unlock the transformative potential of AI agents in biomedical research. Collaborative agentic approaches can lead to groundbreaking advances, ultimately improving human health and well-being.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of NIH R01-HD108794, NSF CAREER 2339524, US DoD FA8702-15-D-0001, and ARPA-H BDF Toolbox Program, and awards from Harvard Data Science Initiative, Amazon Faculty Research, Google Research Scholar Program, AstraZeneca Research, Roche Alliance with Distinguished Scientists, Sanofi iDEA-iTECH Award, Pfizer Research, Chan Zuckerberg Initiative, John and Virginia Kaneb Fellowship award at Harvard Medical School, Biswas Computational Biology Initiative in partnership with the Milken Institute, Harvard Medical School Dean's Innovation Awards for the Use of AI in Research, and Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. A.F. is supported by the Kempner Institute Graduate Fellowship. A.N. is supported by the Herchel Smith-Harvard Undergraduate Science Fellowship, the Yun Family Research Fellows Fund for Revolutionary Thinking, and the Summer Institute in Biomedical Informatics at Harvard Medical School. V.G. is supported by the Medical Research Council, MR/W00710X/1. Y.E. is supported by grant T32 HG002295 from the National Human Genome Research Institute and the NSDEG fellowship. The authors would like to thank Owen Queen, Alejandro Velez-Arce, and Ruth Johnson for their constructive comments on the draft manuscript. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

AUTHOR CONTRIBUTIONS

All authors contributed to the design and writing of the manuscript, helped shape the research, provided critical feedback, and commented on the manuscript and its revisions. M.Z. conceived the study and was in charge of overall direction and planning.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Boiko, D.A., MacKnight, R., Kline, B., and Gomes, G. (2023). Autonomous chemical research with large language models. Nature *624*, 570–578. https://doi.org/10.1038/s41586-023-06792-0.
2. Bran, A.M., Cox, S., Schilter, O., Baldassari, C., White, A.D., and Schwaller, P. (2024). Augmenting large language models with chemistry tools. Nat. Mach. Intell. *6*, 525–535. https://doi.org/10.1038/s42256-024-00832-8. https://openreview.net/forum?id=wdGIL6lx3l.
3. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al. (2023). The rise and potential of large language model based agents: A survey. Preprint at arXiv.
4. Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., and Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. Preprint at arXiv.
5. Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. (2023). Scientific discovery in the age of artificial intelligence. Nature *620*, 47–60. https://doi.org/10.1038/s41586-023-06221-2.
6. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. Preprint at arXiv.
7. Gemini Team, Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Jiahui, Y., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al. (2023). Gemini: a family of highly capable multimodal models. Preprint at arXiv.
8. Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
9. Vemprala, S., Bonatti, R., Bucker, A., and Kapoor, A. (2023). Chatgpt for robotics: design principles and model abilities. Microsoft Auton. Syst. Robot. Res *2*, 20.
10. Significant Gravitas (2023). Autogpt. https://agpt.co.
11. Yao, S., Zhao, J., Dian, Y., Du, N., Shafran, I., Narasimhan, K.R., and Cao, Y. (2023). React: Synergizing reasoning and acting in language models. In The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id=WE_vluYUL-X.
12. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K.R., and Yao, S. (2023). Reflexion: language agents with verbal reinforcement learning. In Thirty-seventh Conference on Neural Information Processing Systems.
13. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., and Wang, C. (2023). Autogen: Enabling next-gen llm applications via multi-agent conversation framework. Preprint at arXiv.
14. Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., and Garg, A. (2023). Progprompt: Generating situated robot task plans using large language models. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 11523–11530. https://doi.org/10.1109/ICRA48891.2023.10161317.
15. Huang, W., Abbeel, P., Pathak, D., and Mordatch, I. (2022). Language models as zero-shot planners: extracting actionable knowledge for embodied agents. In International Conference on Machine Learning (PMLR), pp. 9118–9147.

16. Krenn, M., Pollice, R., Guo, S.Y., Aldeghi, M., Cervera-Lierta, A., Friederich, P., dos Passos Gomes, G., Häse, F., Jinich, A., Nigam, A.K., et al. (2022). On scientific understanding with artificial intelligence. Nat. Rev. Phys. *4*, 761–769. https://doi.org/10.1038/s42254-022-00518-3.

17. Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Gao, C., Huang, Y., Lyu, W., Zhang, Y., Li, X., et al. (2024). Trustllm: Trustworthiness in large language models. Preprint at arXiv.

18. Kotha, S., Springer, J.M., and Raghunathan, A. (2023). Understanding catastrophic forgetting in language models via implicit inference. Preprint at arXiv.

19. Li, H., Moon, J.T., Purkayastha, S., Celi, L.A., Trivedi, H., and Gichoya, J.W. (2023). Ethics of large language models in medicine and medical research. Lancet Digit. Health *5*, e333–e335. https://doi.org/10.1016/S2589-7500(23)00083-3.

20. Goetz, L., Trengove, M., Trotsyuk, A., and Federico, C.A. (2023). Unreliable llm bioethics assistants: Ethical and pedagogical risks. Am. J. Bioeth. *23*, 89–91. https://doi.org/10.1080/15265161.2023.2249843.

21. Kumar, A., Singh, S., Murty, S.V., and Ragupathy, S. (2024). The ethics of interaction: Mitigating security threats in llms. Preprint at arXiv.

22. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., and Nolan, G.P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. Science *308*, 523–529. https://doi.org/10.1126/science.1105809.

23. Rao, R.M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., and Rives, A. (2021). Msa transformer. In International Conference on Machine Learning (PMLR), pp. 8844–8856.

24. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130. https://doi.org/10.1126/science.ade2574.

25. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876. https://doi.org/10.1126/science.abj8754.

26. Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838. https://doi.org/10.1038/nbt.3300.

27. Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., Al Sayed, Z.R., Hill, M.C., Mantineo, H., Brydon, E.M., Zeng, Z., Liu, X.S., et al. (2023). Transfer learning enables predictions in network biology. Nature *618*, 616–624. https://doi.org/10.1038/s41586-023-06139-9.

28. Yu, M.K., Kramer, M., Dutkowski, J., Srivas, R., Licon, K., Kreisberg, J.F., Ng, C.T., Krogan, N., Sharan, R., and Ideker, T. (2016). Translation of Genotype to Phenotype by a Hierarchy of Cell Subsystems. Cell Syst. *2*, 77–88. https://doi.org/10.1016/j.cels.2016.02.003.

29. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer Cell *1*, 203–209. https://doi.org/10.1016/s1535-6108(02)00030-2.

30. Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C.T., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., et al. (2002). Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. *8*, 68–74.

31. Kuenzi, B.M., Park, J., Fong, S.H., Sanchez, K.S., Lee, J., Kreisberg, J.F., Ma, J., and Ideker, T. (2020). Predicting drug response and synergy using a deep learning model of human cancer cells. Cancer Cell *38*, 672–684.e6. https://doi.org/10.1016/j.ccell.2020.09.014.

32. Ren, F., Aliper, A., Chen, J., Zhao, H., Rao, S., Kuppe, C., Ozerov, I.V., Zhang, M., Witte, K., Kruse, C., et al. (2024). A small-molecule tnik inhibitor targets fibrosis in preclinical and clinical models. Published online March 8, 2024. Nat. Biotechnol. https://doi.org/10.1038/s41587-024-02143-0.

33. Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. Cell *180*, 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021.

34. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res. *28*, 235–242. https://doi.org/10.1093/nar/28.1.235.

35. 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65. https://doi.org/10.1038/nature11632.

36. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). Drugbank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. *34*, D668–D672. https://doi.org/10.1093/nar/gkj067.

37. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444. https://doi.org/10.1093/nar/gkab1061.

38. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

39. Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. *30*, 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X.

40. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

41. Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., et al. (2017). The chembl database in 2017. Nucleic Acids Res. *45*, D945–D954. https://doi.org/10.1093/nar/gkw1074.

42. Van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2024). Fast and accurate protein structure search with foldseek. Nat. Biotechnol. *42*, 243–246.

43. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. Siren's song in the ai ocean: a survey on hallucination in large language models. Preprint at arXiv.

44. Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodriques, S.G., and White, A.D. (2023). Paperqa: Retrieval-augmented generative agent for scientific research. Preprint at arXiv.

45. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems *25*.

46. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems *30*.

48. Hernández-García, A., Saxena, N., Jain, M., Liu, C.H., and Bengio, Y. (2024). Multi-fidelity active learning with GFlownets. ICLR. https://openreview.net/forum?id=3QR230r11w.

49. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Gray, A., et al. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems, A.H. Oh, A. Agarwal, D.

Belgrave, and K. Cho, eds. (ICLR). https://openreview.net/forum?id=TG8KACxEON.

50. Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A., Veselov, M.S., Aladinskiy, V.A., Aladinskaya, A.V., Terentiev, V.A., Polykovskiy, D.A., Kuznetsov, M.D., Asadulaev, A., et al. (2019). Deep learning enables rapid identification of potent ddr1 kinase inhibitors. Nat. Biotechnol. *37*, 1038–1040. https://doi.org/10.1038/s41587-019-0224-x.

51. Hie, B.L., and Yang, K.K. (2022). Adaptive machine learning for protein engineering. Curr. Opin. Struct. Biol. *72*, 145–152. https://doi.org/10.1016/j.sbi.2021.11.002. https://linkinghub.elsevier.com/retrieve/pii/S0959440X21001457.

52. Lutz, I.D., Wang, S., Norn, C., Courbet, A., Borst, A.J., Zhao, Y.T., Dosey, A., Cao, L., Xu, J., Leaf, E.M., et al. (2023). Top-down design of protein architectures with reinforcement learning. Science *380*, 266–273. https://doi.org/10.1126/science.adf6591.

53. Bailey, M., Moayedpour, S., Li, R., Corrochano-Navarro, A., Kötter, A., Kogler-Anele, L., Riahi, S., Grebner, C., Hessler, G., Matter, H., et al. (2023). Deep batch active learning for drug discovery. Preprint at bioRxiv.

54. Soleimany, A.P., Amini, A., Goldman, S., Rus, D., Bhatia, S.N., and Coley, C.W. (2021). Evidential deep learning for guided molecular property prediction and discovery. ACS Cent. Sci. *7*, 1356–1367. https://doi.org/10.1021/acscentsci.1c00546.

55. Zhang, J., Cammarata, L., Squires, C., Sapsis, T.P., and Uhler, C. (2023). Active learning for optimal intervention design in causal models. Nat. Mach. Intell. *5*, 1066–1075. https://doi.org/10.1038/s42256-023-00719-0.

56. Yala, A., Mikhael, P.G., Lehman, C., Lin, G., Strand, F., Wan, Y.L., Hughes, K., Satuluru, S., Kim, T., Banerjee, I., et al. (2022). Optimizing risk-based breast cancer screening policies with reinforcement learning. Nat. Med. *28*, 136–143. https://doi.org/10.1038/s41591-021-01599-w. https://www.nature.com/articles/s41591-021-01599-w.

57. Sumers, T., Yao, S., Narasimhan, K., and Griffiths, T. (2024). Cognitive architectures for language agents. Transact Mach Learn Res. https://openreview.net/forum?id=1i6ZCvflQJ.

58. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al. (2024). A survey on large language model based autonomous agents. Front. Comput. Sci. *18*, 1–26. https://doi.org/10.1007/s11704-024-40231-1.

59. Wei, J., Bosma, M., Zhao, V., Guu, K., Adams Wei, Y., Lester, B., Du, N., Dai, A.M., and Le, Q.V. (2022). Finetuned language models are zero-shot learners. In International Conference on Learning Representations. https://openreview.net/forum?id=gEZrGCozdqR.

60. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., and Zou, J. (2023). A visual–language foundation model for pathology image analysis using medical twitter. Nat. Med. *29*, 2307–2316. https://doi.org/10.1038/s41591-023-02504-3.

61. Nori, H., Lee, Y.T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., et al. (2023). Can generalist foundation models outcompete special-purpose tuning? case study in medicine. Preprint at arXiv.

62. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., and Bernstein, M.S. (2023). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pp. 1–22. https://doi.org/10.1145/3586183.3606763.

63. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.Y. (2022). Biogpt: generative pre-trained transformer for biomedical text generation and mining. Brief. Bioinform. *23*, bbac409. https://doi.org/10.1093/bib/bbac409.

64. Jiang, L.Y., Liu, X.C., Nejatian, N.P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H.A., Laufer, I., Punjabi, P., et al. (2023). Health system-scale language models are all-purpose prediction engines. Nature *619*, 357–362.

65. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. Nature *620*, 172–180. https://doi.org/10.1038/s41586-023-06291-2.

66. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Le Hou, K.C., Pfohl, S., Cole-Lewis, H., Neal, D., et al. (2023). Towards expert-level medical question answering with large language models. Preprint at arXiv.

67. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems, *33*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, eds. (Curran Associates, Inc.), pp. 1877–1901.

68. Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. (2023). Voyager: An open-ended embodied agent with large language models. In Intrinsically-Motivated and Open-Ended Learning Workshop @NeurIPS2023. https://openreview.net/forum?id=nfx5IutEed.

69. Fernando, C., Banarse, D.S., Michalewski, H., Osindero, S., and Rocktäschel, T. (2024). Promptbreeder: Self-referential self-improvement via prompt evolution. ICLR. https://openreview.net/forum?id=HKkiX32Zw1.

70. Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q.V., Zhou, D., and Chen, X. (2023). Large language models as optimizers. Preprint at arXiv.

71. LeCun, Y. (2022). A path towards autonomous machine intelligence. Open Rev. *1*, 1–62.

72. Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D.Y., Yang, X., Vodrahalli, K., He, S., Smith, D.S., Yin, Y., et al. (2024). Can large language models provide useful feedback on research papers? a large-scale empirical analysis. NEJM AI *1*, Aloa2400196. https://doi.org/10.1056/Aloa2400196.

73. Justin Chih-yao, C., Saha, S., and Bansal, M. (2024). Reconcile: Round-table conference improves reasoning via consensus among diverse LLMs.. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. *1*, 7066–7085.

74. Sanders, L.M., Scott, R.T., Yang, J.H., Qutub, A.A., Garcia Martin, H., Berrios, D.C., Hastings, J.J.A., Rask, J., Mackintosh, G., Hoarfrost, A.L., et al. (2023). Biological research and self-driving labs in deep space supported by artificial intelligence. Nat. Mach. Intell. *5*, 208–219. https://doi.org/10.1038/s42256-023-00618-4.

75. Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., et al. (2021). Advancing mathematics by guiding human intuition with ai. Nature *600*, 70–74. https://doi.org/10.1038/s41586-021-04086-x.

76. Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G., and Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature *571*, 95–98. https://doi.org/10.1038/s41586-019-1335-8.

77. Jablonka, K.M., Schwaller, P., Ortega-Guerrero, A., and Smit, B. (2024). Leveraging large language models for predictive chemistry. Nat. Mach. Intell. *6*, 161–169. https://doi.org/10.1038/s42256-023-00788-1.

78. Glass, D.J., and Hall, N. (2008). A brief history of the hypothesis. Cell *134*, 378–381. https://doi.org/10.1016/j.cell.2008.07.033.

79. Lim, Y., Tamayo-Orrego, L., Schmid, E., Tarnauskaite, Z., Kochenova, O.V., Gruar, R., Muramatsu, S., Lynch, L., Schlie, A.V., Carroll, P.L., et al. (2023). In silico protein interaction screening uncovers donson's role in replication initiation. Science *381*, eadi3448. https://doi.org/10.1126/science.adi3448.

80. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Le Chi, Q.V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Process. Syst. *35*, 24824–24837.

81. Zhou, J., Zhang, B., Chen, X., Li, H., Xu, X., Chen, S., and Gao, X. (2023). Automated bioinformatics analysis via autoba. Preprint at arXiv.

82. Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., and Gerstein, M. (2023). Medagents: Large language models as collaborators

for zero-shot medical reasoning. Preprint at arXiv. https://doi.org/10.18653/v1/2024.findings-acl.33.

83. Hu, X., Liu, G., Zhao, Y., and Zhang, H. (2023). De novo drug design using reinforcement learning with multiple gpt agents. In Thirty-seventh Conference on Neural Information Processing Systems.

84. Morris, M.R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., and Legg, S. (2023). Levels of agi: Operationalizing progress on the path to agi. Preprint at arXiv.

85. Urbina, F., Lentzos, F., Invernizzi, C., and Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. Nat. Mach. Intell. *4*, 189–191. https://doi.org/10.1038/s42256-022-00465-9. https://www.nature.com/articles/s42256-022-00465-9.

86. Tang, X., Jin, Q., Zhu, K., Yuan, T., Zhang, Y., Zhou, W., Qu, M., Zhao, Y., Tang, J., Zhang, Z., et al. (2024). Prioritizing safeguarding over autonomy: Risks of llm agents for science. Preprint at arXiv.

87. Baker, D., and Church, G. (2024). Protein design meets biosecurity. Science *383*, 349. https://doi.org/10.1126/science.ado1671.

88. Marees, A.T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., and Derks, E.M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int. J. Methods Psychiatr. Res. *27*, e1608. https://doi.org/10.1002/mpr.1608.

89. Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. Nat. Rev. Methods Primers *1*, 59. https://doi.org/10.1038/s43586-021-00056-9.

90. Frueh, F.W. (2010). Real-world clinical effectiveness, regulatory transparency and payer coverage: three ingredients for translating pharmacogenomics into clinical practice. Pharmacogenomics *11*, 657–660. https://doi.org/10.2217/pgs.10.46.

91. Panayiotopoulos, C.P. (2005). The Epilepsies: Seizures, Syndromes and Management (Bladon Medical Publishing).

92. International League Against Epilepsy Consortium on Complex Epilepsies (2023). Gwas meta-analysis of over 29,000 people with epilepsy identifies 26 risk loci and subtype-specific genetic architecture. Nat. Commun. *55*, 1471–1482. https://doi.org/10.1038/s41467-017-02088-w.

93. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779. https://doi.org/10.1371/journal.pmed.1001779.

94. Gamirova, R., Shagimardanova, E., Sato, T., Kannon, T., Gamirova, R., and Tajima, A. (2024). Identification of potential disease-associated variants in idiopathic generalized epilepsy using targeted sequencing. J. Hum. Genet. *69*, 59–67. https://doi.org/10.1038/s10038-023-01208-3.

95. Oliver, K.L., Scheffer, I.E., Bennett, M.F., Grinton, B.E., Bahlo, M., and Berkovic, S.F. (2023). Genes4Epilepsy: An epilepsy gene resource. Epilepsia *64*, 1368–1375.

96. Salowe, R.J., Lee, R., Zenebe-Gete, S., Vaughn, M., Gudiseva, H.V., Pistilli, M., Kikut, A., Becker, E., Collins, D.W., He, J., et al. (2022). Recruitment strategies and lessons learned from a large genetic study of African Americans. PLoS Glob. Public Health *2*, e0000416. https://doi.org/10.1371/journal.pgph.0000416.

97. Aissani, B. (2014). Confounding by linkage disequilibrium. J. Hum. Genet. *59*, 110–115. https://doi.org/10.1038/jhg.2013.130.

98. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The human cell atlas. eLife *6*, e27041. https://doi.org/10.7554/eLife.27041. https://elifesciences.org/articles/27041.

99. Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell *171*, 1437–1452.e17. https://doi.org/10.1016/j.cell.2017.10.049. https://linkinghub.elsevier.com/retrieve/pii/S0092867417313090.

100. Mitchell, D.C., Kuljanin, M., Li, J., Van Vranken, J.G., Bulloch, N., Schweppe, D.K., Huttlin, E.L., and Gygi, S.P. (2023). A proteome-wide atlas of drug mechanism of action. Nat. Biotechnol. *41*, 845–857. https://doi.org/10.1038/s41587-022-01539-0.

101. Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the cancer cell line encyclopedia. Nature *569*, 503–508. https://doi.org/10.1038/s41586-019-1186-3.

102. Chandrasekaran, S.N., Cimini, B.A., Goodale, A., Miller, L., Kost-Alimova, M., Jamali, N., Doench, J.G., Fritchman, B., Skepner, A., Melanson, M., et al. (2022). Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. Nat. Methods *21*, 1114–1121.

103. De Teresa-Trueba, I., Goetz, S.K., Mattausch, A., Stojanovska, F., Zimmerli, C.E., Toro-Nahuelpan, M., Cheng, D.W.C., Tollervey, F., Pape, C., Beck, M., et al. (2023). Convolutional networks for supervised mining of molecular patterns within cellular context. Nat. Methods *20*, 284–294. https://doi.org/10.1038/s41592-022-01746-2. https://www.nature.com/articles/s41592-022-01746-2.

104. Schiøtz, O.H., Kaiser, C.J.O., Klumpe, S., Morado, D.R., Poege, M., Schneider, J., Beck, F., Klebl, D.P., Thompson, C., and Plitzko, J.M. (2023). Serial lift-out: sampling the molecular anatomy of whole organisms. Nat. Methods *21*, 1684–1692.

105. Lundberg, E., and Borner, G.H.H. (2019). Spatial proteomics: a powerful discovery tool for cell biology. Nat. Rev. Mol. Cell Biol. *20*, 285–302. https://doi.org/10.1038/s41580-018-0094-y.

106. Cho, N.H., Cheveralls, K.C., Brunner, A.D., Kim, K., Michaelis, A.C., Raghavan, P., Kobayashi, H., Savy, L., Li, J.Y., Canaj, H., et al. (2022). Opencell: Endogenous tagging for the cartography of human cellular organization. Science *375*, eabi6983. https://doi.org/10.1126/science.abi6983.

107. Johnson, G.T., Agmon, E., Akamatsu, M., Lundberg, E., Lyons, B., Ouyang, W., Quintero-Carmona, O.A., Riel-Mehan, M., Rafelski, S., and Horwitz, R. (2023). Building the next generation of virtual cells to understand cellular biology. Biophys. J. *122*, 3560–3569. https://doi.org/10.1016/j.bpj.2023.04.006. https://linkinghub.elsevier.com/retrieve/pii/S0006349523002369.

108. Li, M.M., Huang, Y., Sumathipala, M., Liang, M.Q., Valdeolivas, A., Ananthakrishnan, A.N., Liao, K., Marbach, D., and Zitnik, M. (2024). Contextual AI models for single-cell protein biology. Nat. Methods *21*, 1546–1557. https://doi.org/10.1038/s41592-024-02341-3.

109. Russell, A.J.C., Weir, J.A., Nadaf, N.M., Shabet, M., Kumar, V., Kambhampati, S., Raichur, R., Marrero, G.J., Liu, S., Balderrama, K.S., et al. (2024). Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. Nature *625*, 101–109. https://doi.org/10.1038/s41586-023-06837-4. https://www.nature.com/articles/s41586-023-06837-4.

110. Wik, L., Nordberg, N., Broberg, J., Björkesten, J., Assarsson, E., Henriksson, S., Grundberg, I., Pettersson, E., Westerberg, C., Liljeroth, E., et al. (2021). Proximity extension assay in combination with next-generation sequencing for high-throughput proteome-wide analysis. Mol. Cell. Proteomics *20*, 100168. https://doi.org/10.1016/j.mcpro.2021.100168. https://linkinghub.elsevier.com/retrieve/pii/S1535947621001407.

111. Liu, Y., DiStasio, M., Su, G., Asashima, H., Enninful, A., Qin, X., Deng, Y., Nam, J., Gao, F., Bordignon, P., et al. (2023). High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial cite-seq. Nat. Biotechnol. *41*, 1405–1409. https://doi.org/10.1038/s41587-023-01676-0. https://www.nature.com/articles/s41587-023-01676-0.

112. Yoshikawa, N., Darvish, K., Vakili, M.G., Garg, A., and Aspuru-Guzik, A. (2023). Digital pipette: open hardware for liquid transfer in self-driving laboratories. Digit. Discov. *2*, 1745–1751. https://doi.org/10.1039/

D3DD00115F. https://pubs.rsc.org/en/content/articlelanding/2023/dd/d3dd00115f.

113. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. Cell *167*, 1853–1866.e17. https://doi.org/10.1016/j.cell.2016.11.038. https://linkinghub.elsevier.com/retrieve/pii/S0092867416316105.

114. Binan, L., Danquah, S., Valakh, V., Simonton, B., Bezney, J., Nehme, R., Cleary, B., and Farhi, S.L. (2023). Simultaneous crispr screening and spatial transcriptomics reveals intracellular, intercellular, and functional transcriptional circuits. Preprint at biorxiv. https://doi.org/10.1101/2023.11.30.569494.

115. Dang, C.V., Reddy, E.P., Shokat, K.M., and Soucek, L. (2017). Drugging the 'undruggable'cancer targets. Nat. Rev. Cancer *17*, 502–508. https://doi.org/10.1038/nrc.2017.36.

116. Lieber, T., Jeedigunta, S.P., Palozzi, J.M., Lehmann, R., and Hurd, T.R. (2019). Mitochondrial fragmentation drives selective removal of deleterious mtdna in the germline. Nature *570*, 380–384. https://doi.org/10.1038/s41586-019-1213-4.

117. Li, G., Al Kader Hammoud, H.A., Itani, H., Khizbullin, D., and Ghanem, B. (2023). Camel: Communicative agents for "mind" exploration of large language model society. In Thirty-seventh Conference on Neural Information Processing Systems.

118. Liu, H., Li, C., Wu, Q., and Lee, Y.J. (2023). Visual instruction tuning. In Thirty-seventh Conference on Neural Information Processing Systems. https://openreview.net/forum?id=w0H2xGHlkw.

119. Chen, J., Poskanzer, K.E., Freeman, M.R., and Monk, K.R. (2020). Live-imaging of astrocyte morphogenesis and function in zebrafish neural circuits. Nat. Neurosci. *23*, 1297–1306. https://doi.org/10.1038/s41593-020-0703-x. https://www.nature.com/articles/s41593-020-0703-x.

120. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Tianhe, Y., et al. (2023). PaLM-e: An embodied multimodal language model. In Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, eds. (PMLR), pp. 8469–8488. https://proceedings.mlr.press/v202/driess23a.html.

121. Li, J., Cai, Z., Vaites, L.P., Shen, N., Mitchell, D.C., Huttlin, E.L., Paulo, J.A., Harry, B.L., and Gygi, S.P. (2021). Proteome-wide mapping of short-lived proteins in human cells. Mol. Cell *81*, 4722–4735.e5. https://doi.org/10.1016/j.molcel.2021.09.015.

122. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In International conference on machine learning (PMLR), pp. 8748–8763.

123. Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2024). MiniGPT-4: enhancing vision-language understanding with advanced large language models. In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=1tZbq88f27.

124. Bavishi, R., Elsen, E., Hawthorne, C., Nye, M., Odena, A., Somani, A., and Taşırlar, S. (2023). Introducing our multimodal models. https://www.adept.ai/blog/fuyu-8b.

125. Kopp, S., and Krämer, N. (2021). Revisiting human-agent communication: the importance of joint co-construction and understanding mental states. Front. Psychol. *12*, 580955. https://doi.org/10.3389/fpsyg.2021.580955.

126. Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., et al. (2022). Inner monologue: Embodied reasoning through planning with language models. In 6th Annual Conference on Robot Learning. https://openreview.net/forum?id=3R3Pz5i0tye.

127. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. Preprint at arXiv.

128. Nascimento, N., Alencar, P., and Cowan, D. (2023). Self-adaptive large language model (llm)-based multiagent systems. In IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C) (IEEE Publications), pp. 104–109. https://doi.org/10.1109/ACSOS-C58168.2023.00048.

129. Lowe, R., Wu, Y.I., Tamar, A., Harb, J., Open, A.I., Abbeel, P., and Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in Neural Information Processing Systems *30*.

130. Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S.K.S., Lin, Z., et al. (2024). MetaGPT: Meta programming for multi-agent collaborative framework. In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=VtmBAGCN7o.

131. Zhang, H., Du, W., Shan, J., Zhou, Q., Du, Y., Tenenbaum, J.B., Shu, T., and Gan, C. (2024). Building cooperative embodied agents modularly with large language models. In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=EnXJfQqy0K.

132. Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., and Shi, S. (2023). Encouraging divergent thinking in large language models through multi-agent debate. Preprint at arXiv.

133. Fu, Y., Peng, H., Khot, T., and Lapata, M. (2023). Improving language model negotiation with self-play and in-context learning from ai feedback. Preprint at arXiv.

134. Mandi, Z., Jain, S., and Song, S. (2023). Roco: Dialectic multi-robot collaboration with large language models. Preprint at arXiv. https://doi.org/10.1109/ICRA57147.2024.10610855.

135. Saha, S., Hase, P., and Bansal, M. (2023). Can language models teach weaker agents? teacher explanations improve students via theory of mind. Preprint at arXiv.

136. Williams, R., Hosseinichimeh, N., Majumdar, A., and Ghaffarzadegan, N. (2023). Epidemic modeling with generative agents. Preprint at arXiv.

137. Park, J.S., Popowski, L., Cai, C., Morris, M.R., Liang, P., and Bernstein, M.S. (2022). Social simulacra: Creating populated prototypes for social computing systems. In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, pp. 1–18. https://doi.org/10.1145/3526113.3545616.

138. Parisi, A., Zhao, Y., and Fiedel, N. (2022). Talm: Tool augmented language models. Preprint at arXiv.

139. Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. (2023). Thirty-seventh Conference on Neural Information Processing Systems. https://openreview.net/forum?id=Yacmpz84TH.

140. Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. (2021). Webgpt: Browser-assisted question-answering with human feedback. Preprint at arXiv.

141. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. (2023). HuggingGPT: Solving AI tasks with chatGPT and its friends in hugging face. In Thirty-seventh Conference on Neural Information Processing Systems. https://openreview.net/forum?id=yHdTscY6Ci.

142. Hu, C., Fu, J., Du, C., Luo, S., Zhao, J., and Zhao, H. (2023). Chatdb: Augmenting llms with databases as their symbolic memory. Preprint at arXiv.

143. Coley, C.W., Thomas, D.A., III, Lummiss, J.A.M., Jaworski, J.N., Breen, C.P., Schultz, V., Hart, T., Fishman, J.S., Rogers, L., Gao, H., et al. (2019). A robotic platform for flow synthesis of organic compounds informed by ai planning. Science *365*, eaax1566. https://doi.org/10.1126/science.aax1566.

144. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al. (2022). Do as i can, not as i say: Grounding language in robotic affordances. Preprint at arXiv.

145. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In International Conference on Machine Learning (PMLR), pp. 8821–8831.

146. Hu, E.J., shen, y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations. https://openreview.net/forum?id=nZeVKeeFYf9.

147. Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu, Z., and Sun, M. (2023). Communicative agents for software development. Preprint at arXiv.

148. Zhou, X., Li, G., and Liu, Z. (2023). Llm as dba. Preprint at arXiv.

149. Zhu, X., Chen, Y., Tian, H., Tao, C., Su, W., Yang, C., Huang, G., Li, B., Lu, L., Wang, X., et al. (2023). Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory. Preprint at arXiv.

150. Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J.M., Tworek, J., Yuan, Q., Tezak, N., Kim, J.W., Hallacy, C., et al. (2022). Text and code embeddings by contrastive pre-training. Preprint at arXiv.

151. Zhong, W., Guo, L., Gao, Q., Ye, H., and Wang, Y. (2024). Memorybank: Enhancing large language models with long-term memory. In Proceedings of the AAAI Conference on Artificial Intelligence, *38*, pp. 19724–19731. https://doi.org/10.1609/aaai.v38i17.29946. https://ojs.aaai.org/index.php/AAAI/article/view/29946.

152. Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. In Thirty-seventh Conference on Neural Information Processing Systems. https://openreview.net/forum?id=OUIFPHEgJU.

153. Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in gpt. In Advances in Neural Information Processing Systems.

154. Zhang, N., Yao, Y., Tian, B., Wang, P., Deng, S., Wang, M., Xi, Z., Mao, S., Zhang, J., Ni, Y., et al. (2024). A comprehensive study of knowledge editing for large language models. Preprint at arXiv.

155. Rana, K., Haviland, J., Garg, S., Abou-Chakra, J., Reid, I., and Suenderhauf, N. (2023). Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In 7th Annual Conference on Robot Learning. https://openreview.net/forum?id=wMpOMO0Ss7a.

156. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org.

157. Li, D., Rawat, A.S., Zaheer, M., Wang, X., Lukasik, M., Veit, A., Felix, Y., and Kumar, S. (2023). Large language models with controllable working memory. In Findings of the Association for Computational Linguistics: ACL, A. Rogers, J. Boyd-Graber, and N. Okazaki, eds. (Association for Computational Linguistics), pp. 1774–1793. https://doi.org/10.18653/v1/2023.findings-acl.112. https://aclanthology.org/2023.findings-acl.112.

158. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems *35*, 22199–22213.

159. Liu, B., Jiang, Y., Zhang, X., Liu, Q., Zhang, S., Biswas, J., and Stone, P. (2023). Llm+ p: Empowering large language models with optimal planning proficiency. Preprint at arXiv.

160. Dagan, G., Keller, F., and Lascarides, A. (2023). Dynamic planning with a llm. Preprint at arXiv.

161. Zhang, Z., Yao, Y., Zhang, A., Tang, X., Ma, X., He, Z., Wang, Y., Gerstein, M., Wang, R., Liu, G., et al. (2023). Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. Preprint at arXiv.

162. Zhong, S., Huang, Z., Gao, S., Wen, W., Lin, L., Zitnik, M., and Zhou, P. (2024). Let's think outside the box: exploring leap-of-thought in large language models with creative humor generation. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (The IEEE Publications) https://doi.org/10.1109/CVPR52733.2024.01258.

163. Sundara Raman, S., Cohen, V., Rosen, E., Idrees, I., Paulius, D., and Tellex, S. (2022). Planning with large language models via corrective re-prompting. In NeurIPS 2022 Foundation Models for Decision Making Workshop.

164. Yao, S., Dian, Y., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., and Narasimhan, K.R. (2023). Tree of thoughts: Deliberate problem solving with large language models. In Thirty-seventh Conference on Neural Information Processing Systems. https://openreview.net/forum?id=5Xc1ecxO1h.

165. Wang, Y., Jiang, Z., Chen, Z., Yang, F., Zhou, Y., Cho, E., Fan, X., Lu, Y., Huang, X., and Yang, Y. (2023). Recmind: Large language model powered agent for recommendation. Preprint at arXiv. https://doi.org/10.18653/v1/2024.findings-naacl.271.

166. Zhou, D., Schärli, N., Le Hou, J.W., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q.V., and Chi, E.H. (2023). Least-to-most prompting enables complex reasoning in large language models. In The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id=WZH7099tgfM.

167. Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, H., Narang, S., Chowdhery, A., and Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id=1PL1NIMMrw.

168. Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., et al. (2024). Graph of Thoughts: Solving elaborate problems with large language models. In Proceedings of the AAAI Conference on Artificial Intelligence, *38*, pp. 17682–17690. https://doi.org/10.1609/aaai.v38i16.29720.

169. Hao, S., Gu, Y., Ma, H., Hong, J.J., Wang, Z., Wang, D.Z., and Hu, Z. (2023). Reasoning with language model is planning with world model. In The 2023 Conference on Empirical Methods in Natural Language Processing https://doi.org/10.18653/v1/2023.emnlp-main.507. https://openreview.net/forum?id=VTWWvYtF1R.

170. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. (2023). Self-refine: Iterative refinement with self-feedback. In Thirty-seventh Conference on Neural Information Processing Systems. https://openreview.net/forum?id=S37hOerQLB.

171. Song, C.H., Wu, J., Washington, C., Sadler, B.M., Chao, W.-L., and Su, Y. (2023). Llm-planner: Few-shot grounded planning for embodied agents with large language models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2998–3009.

172. Chen, X., Lin, M., Schärli, N., and Zhou, D. (2024). Teaching large language models to self-debug. In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=KuPixIqPiq.

173. Wang, R., Zelikman, E., Poesia, G., Pu, Y., Haber, N., and Goodman, N.D. (2023). Hypothesis search: Inductive reasoning with language models. Preprint at arXiv.

174. McCoy, R.T., Yao, S., Friedman, D., Hardy, M., and Griffiths, T.L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. Preprint at arXiv.

175. Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., and Kim, Y. (2023). Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. Preprint at arXiv. https://doi.org/10.18653/v1/2024.naacl-long.102.

176. Nye, M., Andreassen, A.J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. (2021). Show your work: Scratchpads for intermediate computation with language models. Preprint at arXiv.

177. Chen, J., and Mueller, J. (2023). Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. Preprint at arXiv.

178. Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C.D. (2023). Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In The 2023 Conference on Empirical Methods in Natural Language Processing. https://openreview.net/forum?id=g3faCfrwm7.

179. Kuhn, L., Gal, Y., and Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id=VD-AYtP0dve.

180. Shafer, G., and Vovk, V. (2008). A tutorial on conformal prediction. J. Mach. Learn. Res. *9*, 371–421.

181. Shafer, G., and Vovk, V. (2008). A tutorial on conformal prediction. J. Mach. Learn. Res. *9*, 371–421. http://jmlr.org/papers/v9/shafer08a.html.

182. Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. (2023). Discovering language model behaviors with model-written evaluations. In Findings of the Association for Computational Linguistics: ACL, A. Rogers, J. Boyd-Graber, and N. Okazaki, eds. (Association for Computational Linguistics), pp. 13387–13434. https://doi.org/10.18653/v1/2023.findings-acl.847. https://aclanthology.org/2023.findings-acl.847.

183. Qi, X., Zeng, Y., Xie, T., Chen, P.Y., Jia, R., Mittal, P., and Henderson, P. (2024). Fine-tuning aligned language models compromises safety, even when users do not intend to!. In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=hTEGyKf0dZ.

184. Wei, A., Haghtalab, N., and Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? In Thirty-seventh Conference on Neural Information Processing Systems. https://openreview.net/forum?id=jA235JGM09.

185. Bommasani, R., Liang, P., and Lee, T. (2023). Holistic evaluation of language models. Ann. N. Y. Acad. Sci. *1525*, 140–146. https://doi.org/10.1111/nyas.15007.

186. Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., and Scialom, T. (2024). GAIA: a benchmark for general AI assistants. In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=fibxvahvs3.

187. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research. https://openreview.net/forum?id=uyTL5Bvosj.

188. Huang, Q., Vora, J., Liang, P., and Leskovec, J. (2023). Benchmarking large language models as ai research agents. Preprint at arXiv. https://api.semanticscholar.org/CorpusID:263671541.

189. Liu, X., Hao, Y., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., et al. (2024). Agentbench: Evaluating LLMs as agents. In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=zAdUB0aCTQ.

190. Corsello, S.M., Bittker, J.A., Liu, Z., Gould, J., McCarren, P., Hirschman, J.E., Johnston, S.E., Vrcic, A., Wong, B., Khan, M., et al. (2017). The drug repurposing hub: a next-generation drug library and information resource. Nat. Med. *23*, 405–408. https://doi.org/10.1038/nm.4306.

191. Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., Hinsen, K., Larmande, P., Le Bras, Y.L., Lemoine, F., et al. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. Future Gener. Comput. Syst. *75*, 284–298. https://doi.org/10.1016/j.future.2017.01.012.

192. Lamprecht, A.L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E.M., Dominguez Del Angel, V., Van De Sandt, S., Ison, J., Martinez, P.A., et al. (2020). Towards fair principles for research software. Data Sci. *3*, 37–59. https://doi.org/10.3233/DS-190026.

193. Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M.M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Inf. Fusion *50*, 71–91. https://doi.org/10.1016/j.inffus.2018.09.012. https://www.sciencedirect.com/science/article/pii/S1566253518304482.

194. Office of Science And Technology Policy (2023). Ai bill of rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

195. Guha, N., Lawrence, C., Gailmard, L.A., Rodolfa, K., Surani, F., Bommasani, R., Raji, I., Cuéllar, M.F., Honigsberg, C., Liang, P., et al. (2023). Ai regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. George Wash. Law Rev. *11*, 4634443. https://ssrn.com/abstract=4634443.

196. Yang, Z., Raman, S.S., Shah, A., and Tellex, S. (2024). Plug in the safety chip: Enforcing constraints for llm-driven robot agents. In International Conference on Robotics and Automation.

197. Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H.P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. Preprint at arXiv.

198. Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., and Zeng, A. (2023). Code as policies: Language model programs for embodied control. In IEEE International Conference on Robotics and Automation (ICRA) (IEEE Publications), pp. 9493–9500. https://doi.org/10.1109/ICRA48891.2023.10160591.

199. Schwartz, S., Yaeli, A., and Shlomov, S. (2023). Enhancing trust in llm-based ai automation agents: New considerations and future challenges. In International Joint Conference on Artificial Intelligence.