Empowering Biomedical Discovery with AI Agents

Shanghua Gao[1], Ada Fang[1,2,8,+], Yepeng Huang[1,3,+], Valentina Giunchiglia[1,4,+], Ayush Noori[1,5,+], 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, MA, USA*
[9]*Broad Institute of MIT and Harvard, Cambridge, MA, USA*
[10]*Harvard Data Science Initiative, Cambridge, MA, USA*
[+]Co-second authors
[#]Correspondence: marinka@hms.harvard.edu

Summary

We envision "AI scientists" as systems capable of skeptical learning and reasoning that empower biomedical research through collaborative agents that integrate machine learning 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 proficient in a variety of tasks, including self-assessment and planning of discovery workflows. 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 hybrid cell simulation, programmable control of phenotypes, and the design of cellular circuits to the development of new therapies.

1 Introduction

A long-standing ambition for artificial intelligence (AI) in biomedicine is the development of AI systems that could eventually make major scientific discoveries, with the potential to be worthy of a Nobel Prize—fulfilling the Nobel Turing Challenge [1]. While the 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 skeptical learning and reasoning that coordinate large language models (LLMs), machine learning (ML) tools, experimental platforms, or even combinations of them [2–5] (Figure 1).

The complexity of biological problems requires a multistage approach, where decomposing complex questions into simpler tasks is necessary. 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, paving the way toward a future in which a major biomedical discovery is made solely by AI [2, 6]. In the near future, AI agents can accelerate discovery workflows by making them faster and more resource-efficient. Looking ahead, AI agents can enable insights that might not have been possible using ML alone by making predictions across temporal and spatial scales prior to experimental measurements at those scales, and can eventually identify new modes of behavior within biological systems [6].

This vision is possible thanks to advances in LLMs [7–9], multimodal learning, and generative models. Chat-optimized LLMs, such as GPT-4 [10], can incorporate feedback, enabling AI agents to cooperate through conversations with each other and with humans [11]. These conversations can involve agents providing and seeking advice, critique, and validation [12, 13]. 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 [14]. LLMs have also demonstrated the ability to solve complex tasks by breaking them into subtasks [15, 16]. 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 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 [17], so the latter approach aligns well with techniques for next-token prediction within LLMs, while the former does 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 cooperate to solve complex tasks. An AI agent should be skeptical when developing biological hypotheses, capable of characterizing its uncertainty and using that as a driver to acquire and refine its scientific knowledge bases in a way that human scientist partners can trust [18]. 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 and understanding [17]. However, the challenge lies in developing algorithms that allow for this continual learning without suffering from catastrophic forgetting [19], which requires a balance between acquiring new knowledge and retaining existing information. By addressing this challenge, AI agents can power platforms, such as closed-loop synthesis frameworks, autonomous microscopy facilities, and hybrid cell simulators.

Ethical considerations arise from using AI agents [20, 21]. Allowing AI agents to make changes in environments through tools or calls to experimental platforms can be dangerous. Safeguards and protocols need to be in place to prevent harm and negative outcomes [22]. 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 scientists and their reliance on AI agents. The impact of AI agents on biology, coupled with the challenges they pose, underscores the need for responsible implementation. AI agents have the potential to perform research and operational tasks under the oversight of humans [2].

2 Evolving use of data-driven models in biomedical research

Data-driven models have reshaped biomedical research over the past several decades through the development of databases, search engines, machine learning, and interactive and foundation learning models (Figure 2). These models have advanced modeling of proteins [23–27], genes [28], phenotypes [29], clinical outcomes [30–32], and chemical compounds [33, 34] through mining of biomedical data.

Databases and search engines. In biological research, databases (DBs) [35–37] aggregate knowledge from experiments and studies, offering searchable repositories containing standardized biological data vocabularies. An example of such a database is the AlphaFold Protein Structure DB [38], which includes more than 200 million protein structures predicted by AlphaFold [39]. Molecular search engines retrieve information from these databases [40–42]. FoldSeek [43] retrieves protein structures from the AlphaFold DB by translating query structures into 3D interaction alphabet sequences and using pretrained substitution matrices.

Distinct from search engines, designed for retrieving information based on specific queries, AI agents are capable of reasoning to formulate search queries and subsequently acquire information. Curated databases offer structured and factual information, aiding in reducing the risks associated with misinformation potentially generated by agent hallucinations [44]. A notable feature of these agents is their ability to dynamically retrieve information when needed and to create and reflect upon the obtained passages. This reflection process renders the agent controllable during inference, allowing for customization of its actions to meet the requirements of tasks beyond what is possible using search engines and database queries.

Machine learning models. Beyond information retrieval, ML models excel in identifying patterns and assimilating latent knowledge to generalize predictions about novel data [45–47]. An example is AlphaFold [39] protein folding model that uses multi-sequence alignment with a deep learning model to predict the 3D protein structure from amino acid sequences, achieving near-atom level accuracy. AI agents represent an evolution in ML models, building on the foundations of successes such as the transformer architecture [48] and generative pretraining [9]. These agents have reasoning and interactive capabilities that 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 [49] and reinforcement learning [50], represents a further advancement in ML models. This method enhances the adaptability and efficiency of ML techniques by incorporating exploration and feedback mechanisms. Active learning focuses on efficiently training models with limited labeled data, proving particularly advantageous in biological learning where data may be insufficient. 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 applied to various fields, such as small molecule design [51], protein design [52, 53], drug discovery [54, 55], perturbation experiment design [56], and cancer screening [57]. For instance, GENTRL [51] uses reinforcement learning to navigate the chemical space and find candidate chemical compounds that can act against a biological target.

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) [50] uses a "reward model" to train and fine-tune an LLM-based agent with direct human feedback to understand human instruction naturally.

AI agents. Looking ahead, AI agents have advanced capabilities, including proactive information acquisition through perception modules, interaction with tools, reasoning, and the capability to engage with and learn 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 (Section 5). 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 [2, 3].

3 Types of biomedical AI agents

The prevailing approach to building agents nowadays is to use LLMs, where a single LLM is programmed to perform various roles (Section 3.1). 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 (Section 3.2). Given that much of biomedical research is not text-based, such agents have broader applicability to biomedicine.

3.1 Large language model 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 [58, 59] (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 [60] can be used for the former by training the LLM to follow human instruction through the use of prompt examples, including dialogues that incorporate biological reasoning [61]. 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 [62] and performing tasks previously not possible [63]. Early results in clinical medicine question-answering suggest that assigning specific roles to GPT-4 [62] can achieve better performance than using domain-specialized LLMs like BioGPT [64], NYUTron [65], and Med-PaLM [66, 67].

We envision three approaches for assigning roles to biological AI agents: domain-specific fine-tuning, in-context learning, and automatic generation of optimized roles. The first approach involves instruction-tuning an LLM across a large number of biological tasks to ground the LLM in the biological domain, followed by RLHF to ensure that the tuned LLM performs tasks aligned with the goals, wants, and needs of scientists. The second approach uses in-context learning of LLMs [68] 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 [63, 69]. Both strategies require biologists to carefully gather task-specific data or craft prompts. However, as roles defined by humans may not always direct agents as intended, there has been a movement towards allowing LLM-based agents more autonomy in role specification. This paradigm shift in role definition enables agents to autonomously generate and refine role prompts, engaging in self-directed learning and role identification. For instance, Fernando et al. [70] demonstrate an agent's ability to evolve and tailor its prompts in reaction to user inputs. Similarly, Yang et al. [71] investigate the application of LLM as an optimizer to enhance prompt refinement and optimization for improved performance in assigned roles. Through this self-referential learning framework, agents transition from task executors to entities capable of more autonomous learning.

The agent system, comprising a single LLM prompted to adopt various roles, has proven to be an effective auxiliary in scientific research. Studies suggest that agents allocated specific roles exhibit enhanced capabilities compared to either sequentially querying a single LLM or employing a single tool repetitively. A case in point is Coscientist [2], 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 support the use of tools such as web and documentation search, code execution via Python API, and even symbolic lab language (SLL) [2]. 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 the 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.

3.2 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 [72]. 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 schemes for multi-agent systems.

Brainstorming agents (Figure 3b). Brainstorming research ideas with multiple agents 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 and resources to form a collective idea pool. This pool can then be distilled and examined more thoroughly. To foster divergent thinking and creativity, it can be beneficial for agents to specialize in areas of biology, for instance, microglia biology, neuronal degeneration, and neuroinflammation in the case of a multi-agent system for Alzheimer's.

Expert consultation agents (Figure 3c). Expert consultation 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. Liang et al. [73] illustrate that LLMs can function like human expert reviewers, offering scientific critiques on research manuscripts. 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 (Figure 3d). In a research debate, 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.

Round table discussion agents (Figure 3e). Round-table discussions 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. For instance, Reconcile [74] implements a collaborative reasoning scheme among LLM agents through successive rounds of dialogue. 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 (Figure 3f). The self-driving laboratory 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 [75]. 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 [76]. Ideally, hypothesis agents are creative and skeptical when developing biological hypotheses that extrapolate indirectly from the existing body of knowledge [17]. Experimental agents steer operational agents that use a combination of in silico approaches and physical platforms to execute experiments. Following this, 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.

4 Levels of autonomy in AI agents

AI agents, when integrated with experimental platforms, can operate at varying levels of autonomy tailored to the diverse requirements across biomedical fields. We classify these AI agents into four levels according to their proficiency in three areas of discovery: Hypothesis, Experiment, 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**, using ML models as tools. This level aligns with the prevailing approach of using interactive and foundation learning models (Section 2). At this level, ML models do not independently formulate testable and falsifiable statements [77] 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 [78]. 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 multi-modal data to execute these tasks. For instance, ChemCrow [3] combines chain-of-thought reasoning [79] with ML tools to support tasks in organic chemistry, identifying and summarizing literature to inform experiments. In another example, AutoBa [80] 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 [3, 80–82] formulate simple hypotheses inferred from existing knowledge and utilize a limited set of tools, lacking the capacity for a major scientific discovery 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. 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.

The levels of autonomy described for artificial general intelligence (AGI) agents in scientific contexts, particularly in biology, deviate from existing taxonomies that focus on general human-AI interaction separate from the collaborative dynamics between scientists and AI. Existing taxonomies of autonomy consider solely the balance of responsibilities between AI agents and humans—with no consideration of biomedical discovery—and focus on developing AGI to surpass human performance across varying skill levels [83].

As the level of autonomy increases, so does the potential for misuse and the risk of scientists developing an overreliance on agents. While agents have the potential to enhance scientific integrity, there are concerns regarding their use in identifying hazardous substances or controlled substances [84]. Responsible development of agents requires developing preventive measures; further discussion on risks is in [85, 86] and Section 6. The responsible deployment of agents must account for the risk of overreliance, 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 subjected to the same scrutiny as scientists, including reproducibility and rigorous peer review of agentic research.

4.1 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 raw 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 [87], (2) estimate ungenotyped SNPs through imputation, and (3) conduct the appropriate statistical analyses to identify relevant SNPs, taking into account the false discovery rate [88]. 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 then executes tasks on its own in order 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 [89]. 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 [90]—a Level 2 agent would synthesize genetic information from GWAS meta-analyses [91], such as the UK Biobank [92], targeted sequencing studies [93], and knowledge bases like Genes4Epilepsy [94]. 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 [95]. Once data are collected, the agents innovate statistical methods to identify causal variants from genotypic data amidst confounders such as linkage disequilibrium [96] 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.

4.2 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 [97–99], cell viability [100], morphology [101, 102], cryo-electron microscopy and tomography [103, 104], and multiplexed spatial proteomics [105, 106]. This proliferation of data has spurred interest in in silico cell modeling [107] with the overall goal of identifying new mechanistic understanding and adaptive cellular function.

ML tools have been instrumental in analyzing data across these cellular modalities, albeit as Level 0 agents these tools lack autonomous research capabilities. At Level 1, agents integrate specialized Level 0 models designed for narrow tasks such as cell type annotation to assist in hypothesis testing and answering scientific questions. These agents actively engage in research, synthesizing literature and predicting cellular responses using integrated models. For example, to investigate a new compound's mechanism of action (MoA) [99], a Level 1 agent predicts its effects in various cellular contexts [108]. These predictions inform experimental design, such as transcriptomic and proteomic screening. The agent also retrieves and refines experimental protocols for execution on platforms [109]. After data generation, the agent applies predefined bioinformatics pipelines to extract information from multi-modal data before scientists interpret the results.

Level 2 agents not only execute predefined tasks but also generate hypotheses on cellular functions and compound mechanisms. They autonomously refine tasks to support scientific reasoning, enabling more efficient exploration of complex phenotypes like drug resistance through combinatorial perturbations [110]. By managing the experimental cycle and continuously updating their in silico tools, Level 2 agents aim to minimize experimental redundancy and focus on key variables of resistance. However, efficient probing of the intractable combinatorial space of genes and contexts is still a considerable challenge. In light of this, level 2 agents would propose and optimize experiments for a comprehensive yet cost-effective evaluation of hypotheses through knowledge bases, managing the entire experimental cycle from protocol retrieval and improvement to execution, based on iterative feedback from scientists [56]. Automatic data analysis occurs at every experimental stage, with the agent refining hypotheses, adjusting plans, and continuously updating its ML tools based on a synthesis of predictive content, uncertainty, and newly acquired data.

Level 3 agent systems will implement hybrid cell models as digital-experimental simulators of cellular responses to any type of perturbation. Agents in the hybrid cell model system combine AI tools (digital agents) with high-throughput platforms (experimental agents). Digital agents, such as LLM-based agents, autonomously identify knowledge gaps through literature synthesis and handle any perturbagen (extrinsic events such as gene knockouts, compounds, cell-cell interactions; intrinsic events such as cell cycle) based on criteria such as data volume, biological relevance, and clinical needs. Experimental agents generate high-throughput transcriptomics [111] and proteomics measurements [112, 113] with a spatial resolution for perturbations, perturbation screens including in vivo sequencing, and mass spectrometry platforms [113], as well as barcoded spatial proteomics [112, 114]. Optimization of experimental protocols [101, 115, 116] shifts the role of scientists from performing minute operational tasks to managing hybrid cell models.

4.3 Illustration of AI agents in chemical biology

A major focus of chemical biology scientists is on molecular interactions within cells to manipulate biological systems at molecular and cellular levels. An ideal AI agent would be able to 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, progress toward an AI agent proficient in the scientific method within this field remains limited. Existing predictive models for protein structure prediction [39], molecular docking [117], and generative models for protein design [118, 119] are ML tools that can be wrapped within agents through 'function calling' and APIs (Section 5.2). Such agents would be categorized as Level 0, where a scientist oversees all activities. A Level 1 agent, capable of studying a specific protein target, integrates ML tools, such as AlphaFold for structure prediction [39] and neural networks for screening chemical libraries [120] to find candidate chemical compounds that might bind the target. This agent would possess elementary reasoning abilities by probing relevant literature and designing experiments within a specific domain [2], although it would require human guidance for complex inquiries and error correction. The scientist will need to guide the agent to assist with more complex questions and for error checking. For example, a Level 1 agent tasked with binder design for a well-known target would design very similar drug derivatives to existing binders, for which the scientist would have to guide the agent with many different drug scaffolds. However, for a more challenging target, it would provide nonsensical binders.

A Level 2 agent would surpass Level 1 by identifying novel research avenues, such as designing new binders by leveraging trends in related targets; a method demonstrated in materials science for discovering new materials [121]. This agent would employ advanced reasoning extracting literature with more abstract connections to the research task, such as identifying scaffolds that bind to similar pockets and adapting them for the target. In addition, the agent has broader capabilities in planning and designing experiments than a Level 1 agent with expertise in more domains such as retro-synthesis, crystallography, bioassays, and directing robotic arms to execute these experiments. It can also reason from the experiments' results to identify novel discoveries and propose follow-up experiments. Thus, the agent demonstrates a more robust understanding of chemical biology across hypotheses, experiments, and reasoning. The scientist collaborates with the agent by providing feedback on the quality of designed inhibitors and guiding the agent's decisions.

A Level 3 agent is capable of studying all types of molecular interactions in a cell. This agent would work alongside human scientists to explore research questions, proposing de novo binders for an undruggable target [122] or a poorly studied target. Unlike the Level 2 agent's use of well-established protocols, a Level 3 agent unlocks experimental capabilities at spatial and temporal scales that are not currently accessible to experimentation, or design of in situ experiments to study molecular interactions. We can envision the agent proposing candidate molecules, synthesizing molecules with more complex pathways, and designing and executing assays to test efficacy. Finally, it contextualizes results with literature to understand what chemical interactions would occur and why they would be prevalent.

5 Roadmap for building AI agents

AI agents can be obtained through LLM-based systems augmented with different modules [4, 58, 59] that implement functionalities and accomplish tasks detailed in Sections 3 and 4. Here, we describe these modules (Figure 4), focusing on perception, interaction, memory, and reasoning that are necessary for AI agents to engage with human and experimental environments and to make decisions. Interactions between an 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 one. Finally, Reasoning modules process information and execute action plans. Figure 5e illustrates an agent system that uses perception, interaction, memory, and reasoning modules for research in cell biology [123].

5.1 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 receive feedback from multiple sources: scientists [50], the environment [63], and other AI agents [14, 124]. This requires accommodating a diverse array of modalities. These include: text descriptions [7]; images from light and (cryo-)electron microscopy to assess cellular processes across many conditions simultaneously [103, 104, 125]; videos from live imaging to assess developmental processes or animal behaviors across time [126–128]; longitudinal biosensor readouts and genomics profiles of cells [129]; mass spectrometry-based proteomics to decipher protein homeostasis [25, 130]; and miniaturized platforms for conducting biochemical assays and 3D culture systems that mimic the physiological context of organ systems [109].

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 using multi-modal perception modules, where agents process multi-modal data streams from the environment or align multi-modal 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 [50] that it is now possible to build interfaces to agent systems that are entirely based on natural language with limited misinterpretations. The 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 [63, 69]. Combining the history trace of these interactions with retrieval-augmented generation (RAG), it will be possible to develop personalized discovery workflows tailored to individual scientists.

Multi-modal perception modules. Agents align LLMs with other data types to consider data modalities beyond natural language. 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 [10]. 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 [131] for visual information processing. These representations are then aligned with LLM textual representations [125, 132–135] through instruction tuning [125], enabling agents powered by LLMs to perceive and interpret multi-modal data. An alternative to alignment involves allowing the agents to receive input expressed in different modalities [8, 136]. For instance, Fuyu [136] uses a decoder-only transformer architecture to process image patches and text tokens jointly. Similarly, Gemini [8] 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 (Section 5.2) and reasoning (Section 5.4) follow to process the inputs and interact externally.

5.2 Interaction modules

Beyond conversational modules, in biological research, scientists use ML-based and other tools, 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 [137, 138], multi-agent interaction for collaboration among agents, and tool-use action to access ML tools and experimental platforms.

These interaction abilities of LLMs, when combined with interactive 'function calling' (i.e., LLM requesting for tasks to be completed), can act as an intermediary between scientists and the agent's interface, as well as between scientists and various functional items (e.g., tools, other agents, humans). This allows scientists not to search where and how to accomplish tasks but to simply express their intentions in their language. At the same time, the advantages of functional items are fully preserved because agents can interact with tools and use them to provide feedback.

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 [50] 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 [139], 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 [11, 138], using human insight as a directive for carrying out complex tasks [14]. For instance, Inner Monologue [138] employs human feedback to discern user preferences or interpret ambiguous requests in an embodied context. In AutoGPT [11], humans formulate tasks and score solutions returned by agents, and AutoGen [14] 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 of each other. 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 [5, 14, 74, 124, 140]. An instance of this is Generative agents [63], which create interactive environments where agents mimic human behavior and interact using natural language. Different strategies are used for multi-agent interaction, including cooperation [141–143] and negotiation [74, 144, 145]. For example, MetaGPT [142] 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 [81, 146]. MedAgent [81] leverages the expertise of multiple medical AI agents for medical reasoning. Similarly, RoCo [146] 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 [147]. These interactions also enable the creation of simulations for a variety of environments, ranging from public health scenarios [148] to human social behaviors [63, 149], 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 [150].

Commonly used tools are application APIs [151], search engines [152], ML models [153], knowledge databases [154], and robotic machinery for physical tasks [10, 155, 156]. Different Level 1 agent systems have been developed that can interact with one or more types of tools. ChemCrow [3] leverages chemical tools and search engines to address chemical challenges. WebGPT [152] can conduct searches and navigate web browsing environments. SayCan [156] controls a robot in the physical world using an LLM to complete tasks. To invoke these tools, AI agents generate commands in specific formats [151, 153, 154] or query pre-trained control models to execute actions [156, 157]. To develop these capabilities, agents can use in-context learning [153] or fine-tuning with tool-use demonstrations [151], where the latter represents a more sophisticated approach.

In the case of in-context learning, it is necessary to include system abilities in the prompt so agent systems can use 'function calling' to query tools. For example, HuggingGPT [153] 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 [151] 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.

5.3 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 [9, 158], or external, maintained in auxiliary knowledge bases [159, 160]. Internal memory is directly used for accomplishing zero-shot tasks [7, 8] while accessing external memory requires actions by the agent to fetch and integrate data into short-term memory for immediate use [161, 162]. For instance, ChatDB [154] uses an external database for memory storage, and MemoryBank [163] encodes memory segments into embeddings for later retrieval. Agents can query knowledge banks, such as a GWAS database 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 [158, 164], interactive learning [50], and model editing [165]. These strategies must be effective for large models [164] and avoid the loss of previously learned information [166]. On the other hand, updating external memory is more straightforward, involving modifications to the knowledge base [154, 163]. 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 [156, 167] or via latent embeddings [125, 132] in LLMs. For chatbots, previous conversations are kept as text prompts, supporting multiple rounds of dialogue [50, 168]. The text-based approach lays the groundwork for communication in multi-agent [74, 145] and agent-human scenarios [11, 14]. In embodied AI agents, environmental feedback [156, 167] is captured in textual format, acting as a short-term memory that aids reasoning. Following perception, multi-modal inputs are converted into latent embeddings, which function as short-term memory. LLaVA [125] uses latent embeddings generated by visual encoders to retain visual information. Short-term memory allows agents to temporarily acquire skills, such as tool usage [151, 153], store information about recent states of a biological system [167, 168], and keep track of outcomes from earlier reasoning efforts [12]. 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 [169]. 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 homologous protein to inform the initial inhibitor design for the given protein.

5.4 Reasoning modules

Biological research involves a multidisciplinary and multistage process that integrates the expertise of scientists from various disciplines. These scientists formulate hypotheses, design experimental setups based on these hypotheses, interpret the results, and plan the next steps. Central to this process is human reasoning, an ability that conventional ML models find challenging to replicate. The integration of reasoning capabilities in AI agents holds the potential to expedite biological research by assisting in several of these critical steps. Reasoning improves agents' capabilities to plan experiments, make decisions on biological hypotheses, and re-solve competing candidate biological mechanisms. Reasoning modules can be implemented using direct prompting [170] and few-shot in-context learning [79]. Additionally, agents can use planner models [171, 172] and action models [167]. We classify reasoning modules into two categories: direct reasoning and reasoning with feedback. The classification depends on whether the agent adjusts its 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 [173]. For instance, chain-of-thought (CoT) reasoning allows agents to reason step-by-step either by using in-context examples [79] or by applying a zero-shot prompt like "Let's think step-by-step" [170]. Leap-of-thought [174] encourages the model to use creative rather than logical reasoning. Although single-path reasoning matches well with certain situations [175], its ability to adjust to different conditions is limited.

Conversely, multi-path reasoning examines several paths before consolidating them into a final plan [176, 177], allowing for a more thorough planning process that accounts for different scenarios. For example, Least-to-Most prompting [178] breaks down tasks into subproblems solved sequentially. Self-consistent CoT [179] chooses the most consistent answer from a set of CoT answers. Tree-of-thoughts [176] 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 [180] further develops reasoning paths into a graph structure for complex reasoning. To identify the optimal path, methods such as voting strategies [179], Monte Carlo tree search [181], and breadth/depth-first search algorithms [176] are used. Through direct reasoning, agents can generate multiple threads of thought that could consider the best pathways, protein targets, and experiments that can be run 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 [12, 69, 161]. This feedback may include agent-human interaction (Section 5.2) and responses from agents, which can be complementary biological assays quantifying downstream effects of target molecules [182]. In each reasoning cycle, React [12] incorporates insights from previous actions to refine its thought process and inform future actions. LLM-Planner [183] dynamically adjusts plans based on new observations in an embodied environment. Inner Monologue [138] uses both passive and active scene descriptions and feedback from recent actions to guide future actions. Voyager [69] 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 [182, 184]. Techniques like self-refine [182] revise action outputs based on the LLM evaluation, the self-check [182] mechanism allows the agent to review and adjust its reasoning, and reflection [13] 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).

6 Challenges

The perspective outlines key steps to implement AI agents in biomedical research and highlights areas that can benefit from agentic AI. Challenges remain and may, in some cases, be amplified when multi-agent systems become available. Below are technical obstacles and challenges to overcome (Figure 6).

6.1 Robustness and reliability

A barrier facing the deployment of agent systems – specifically those categorized within Levels 2 and 3 as discussed in Section 2 – 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 are unaware of their knowledge gaps) and high sensitivity to the precise formulation of queries, particularly in the context of LLM-based agents.

Such behavior has been traced to the pretraining phase of these models. By contrasting the predicted word sequence and the sequence present in the training data, the autoregressive loss function influences the subsequent model performance with the probability distribution of its inputs, the sequence of generated outputs, and the frequency of tasks it encounters during training [185]. As a result, model performance degrades on task variants that deviate from the assumptions made during training [186].

Sensitivity to input and task probability also offers a potential explanation for the widely observed success of various prompting techniques [79, 176, 187] (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 [129], limitations to understanding physical interactions and long-horizon planning remain. The complexities of training such multi-modal systems, the need for large datasets to cover the range of embodied tasks and environments, and the computational demands of processing multi-modal inputs all remain open questions [8]. Deployment faces challenges from false negatives causing repeated attempts and eventual stalling of the embodied agent [138]. Hence, it is necessary to verify the agent action plan before execution.

Uncertainty quantification can trigger fall-back safety measures like early termination, pre-defined 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 [48, 188]. Techniques such as various forms of prompting, *e.g.*, [179, 189, 190] estimate uncertainty based on the model's predictive distribution, p(output|input), which may itself be subject to bias ([185], Section 3.3); furthermore, it does not consider the distribution of model parameters consistent with the observed training data and marginalizes over its predictions [191]. While conformal prediction [192] 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.

An overarching 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 [193]. Furthermore, fine-tuning existing models with new data can compromise their original alignment, challenging the integrity of the AI agent's intended purpose [194]. Jailbreak attacks can similarly affect post-deployment, highlighting the need for rigorous evaluation [195].

6.2 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 requires transparent change logs and version control for agents, akin to practice in software development.

Evaluation frameworks consider either a holistic evaluation of LLMs [196, 197] or weak spots such as task framing [198, 199], long temporal dependencies, invalid formatting or refusal to follow instructions [200]. A caveat of such methods 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.

6.3 Dataset generation

As laid out, the vision of advanced agents in biology 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 complex systems for biology 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 [201]). 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, multi-modal data requires 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.

6.4 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 [202]. 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 [203].

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 both unintended harmful consequences and malicious intent. Tang et al. [85] describe a safeguarding framework that includes training, licensing, and mandatory safety and ethical compliance checks for agents.

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. 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.

6.5 Risks and safeguards

Developing autonomous experiments and deployment 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 space of ethical considerations relevant to AI agents is too broad to canvass comprehensively here, but this section highlights a few 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.

LLM agents leverage the world knowledge and general reasoning abilities of LLMs obtained during pretraining for solving robotics and planning tasks. However, while considerable effort has been made to teach the robots the "dos," the "don'ts" received less attention. In agent systems, teaching the robots the "don'ts" is crucial: conveying instructions about prohibited actions, assessing the robot's comprehension of these restrictions, and ensuring compliance [204]. For LLM agents, Yang et al. [204] developed a plug-in safety chip, a queryable safety constraint module that translates natural language constraints into formal safety constraints that the robot adheres to. Experiments with robots highlight the potential for integrating formal methods with LLMs for robotic control.

LLMs trained in code completion can write Python programs from docstrings [205]. Given natural language commands, these code-writing LLMs can be re-purposed 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. Writing robot policy code entails using LLMs trained on code completion to write the code from natural language commands [206]. However, the robot policy code is less reliable for enforcing safety constraints than verifiable safe operations that satisfy standards such as 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. They also struggle with understanding physical constraints of the environment in which robots operate. Additionally, generalizing plans across robotic embodiments is brittle with current LLMs. This implies a limitation in what can be generalized, particularly without extensive data collection and adaptability of the generated policies.

7 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 skeptical 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 creates a path to achieve this vision once focused on preventing and learning from mistakes. Building trustworthy sandboxes [207], where AI agents can fail and learn from their mistakes, is one way to achieve this. This involves developing AI agents that perform tasks and consider the boundary of their generalization ability, fostering natural and artificial intelligence.

Declaration of interests The authors declare no competing interests.

Acknowledgments We gratefully acknowledge the support of NIH R01-HD108794, NSF CAREER 2339524, US DoD FA8702-15-D-0001, 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, Aligning Science Across Parkinson's (ASAP) Initiative, Biswas Computational Biology Initiative in partnership with the Milken Institute, and Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. A.F. is supported by the Kempner Institute Graduate Fellowship. 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.

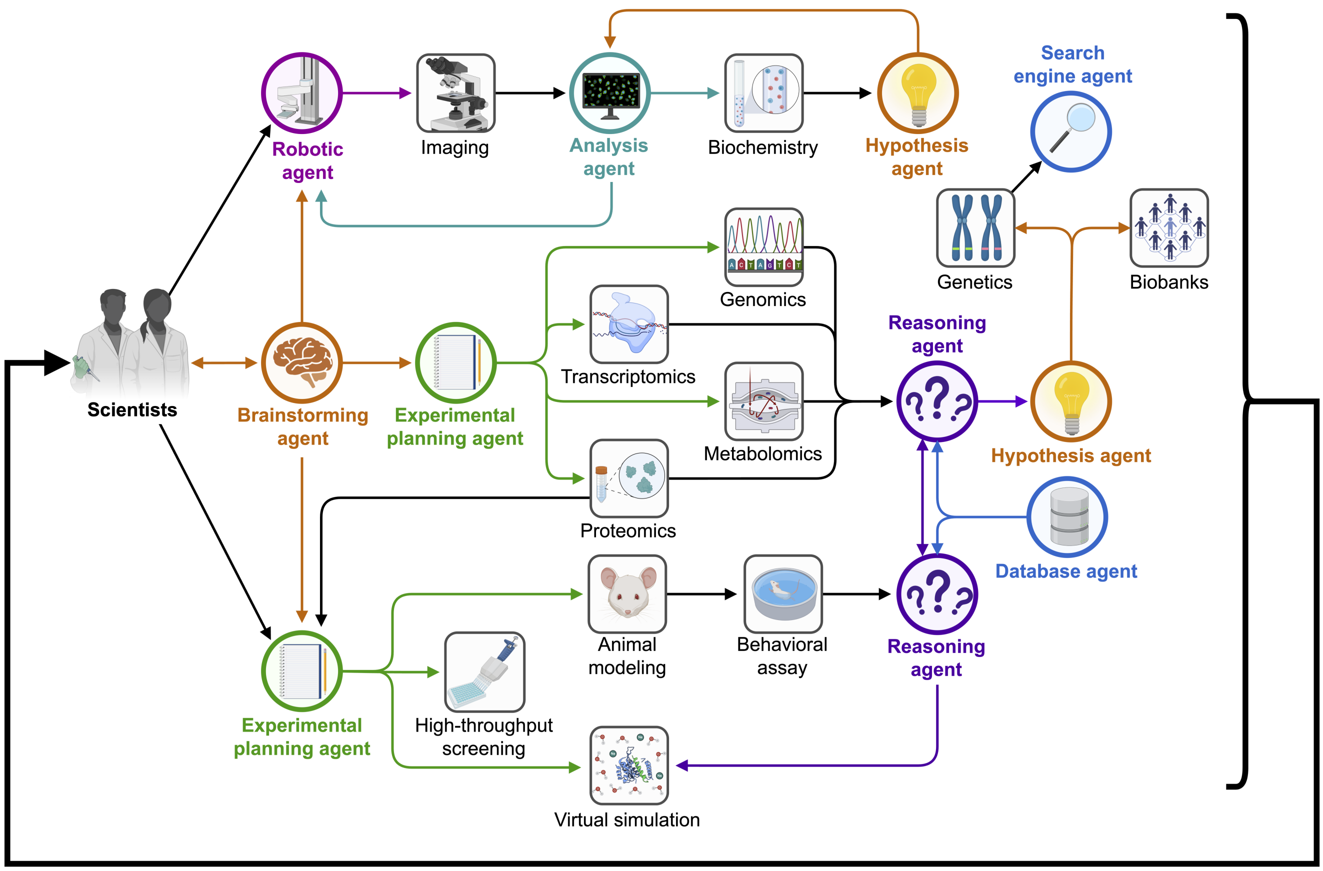

Figure 1: Empowering biomedical research with AI agents. AI agents pave the way for "AI scientists" capable of skeptical learning and reasoning. These multi-agent systems consist of agents based on conversable large language models (LLMs) and can coordinate machine learning (ML) tools, experimental platforms, humans, or even combinations of them. Robotic agent, AI agent that operates robotic hardware for physical experiments; Database agent, AI agent that can information in databases via 'function calling' and APIs; Reasoning agent, AI agent capable of direct reasoning and reasoning with feedback; Hypothesis agent, AI agent that is creative and skeptical 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; Experimental planning agent, AI agent that optimizes an experimental protocol for execution.

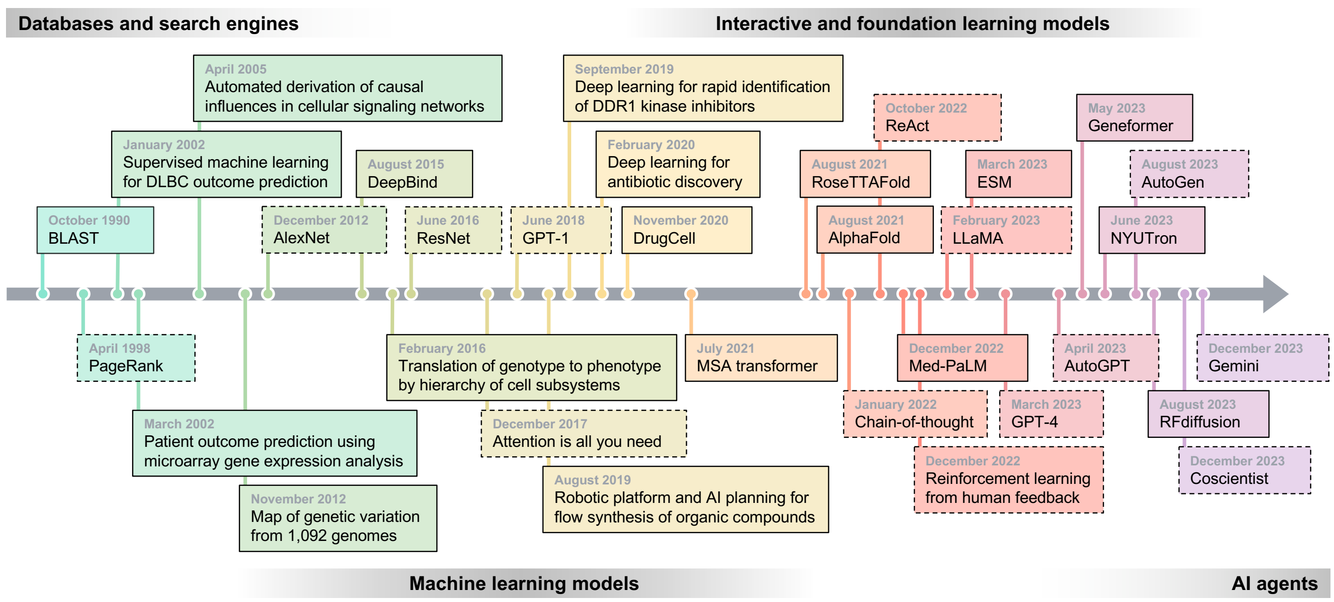

Figure 2: Evolving use of data-driven models in research. Data-driven approaches, from databases and search engines, machine learning, and interactive learning models to advanced agent systems (Section 2), have reshaped biomedical research throughout the last several decades.

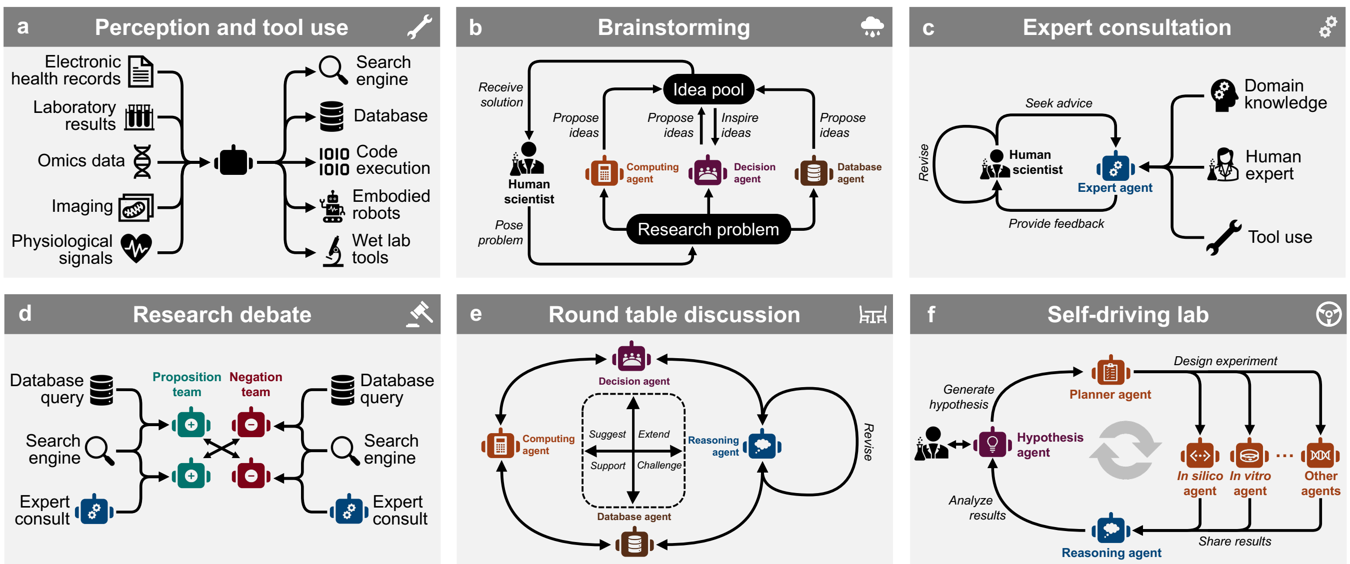

Figure 3: Diverse configurations of AI agents in biology – from an LLM-based AI agent to a multi-agent system with AI models, tools, and integrated physical devices. a. By programming an LLM with the role, one LLM-based agent, equipped with memory and reasoning abilities, performs multi-modal perception 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 schemes, 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 skeptical learning and reasoning to generate hypotheses; Planner agent, AI agent that devises plans for future actions; In silico/vitro agent, AI agent that uses tools in silico or in vitro environment.

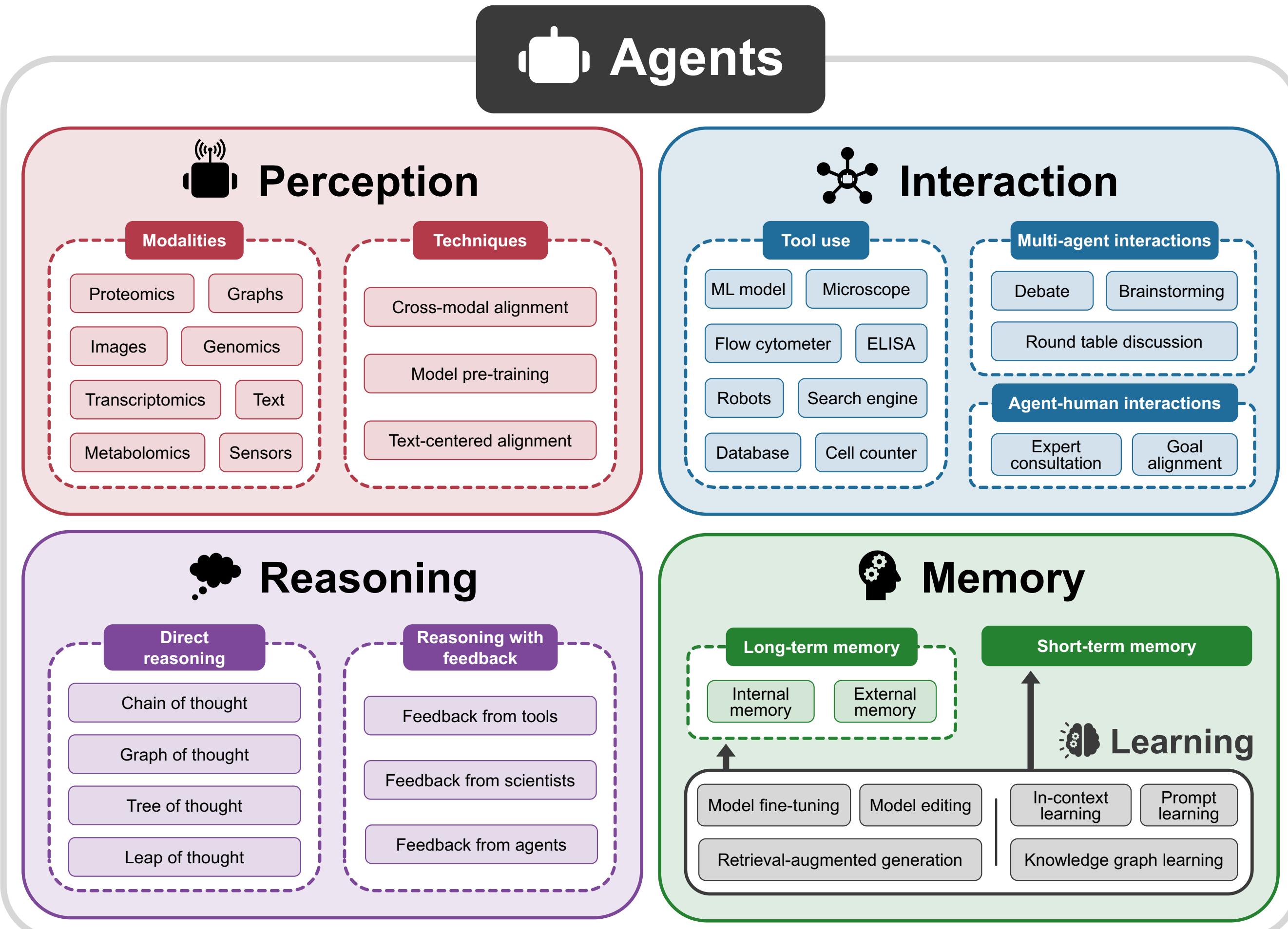

Figure 4: AI agents consist of four key modules: perception, interaction, reasoning, and memory modules. Perception interprets multi-modal environmental data. Interaction facilitates engagement with the environment, encompassing human-agent interactions, multi-agent interactions, and tool use. Memory 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.

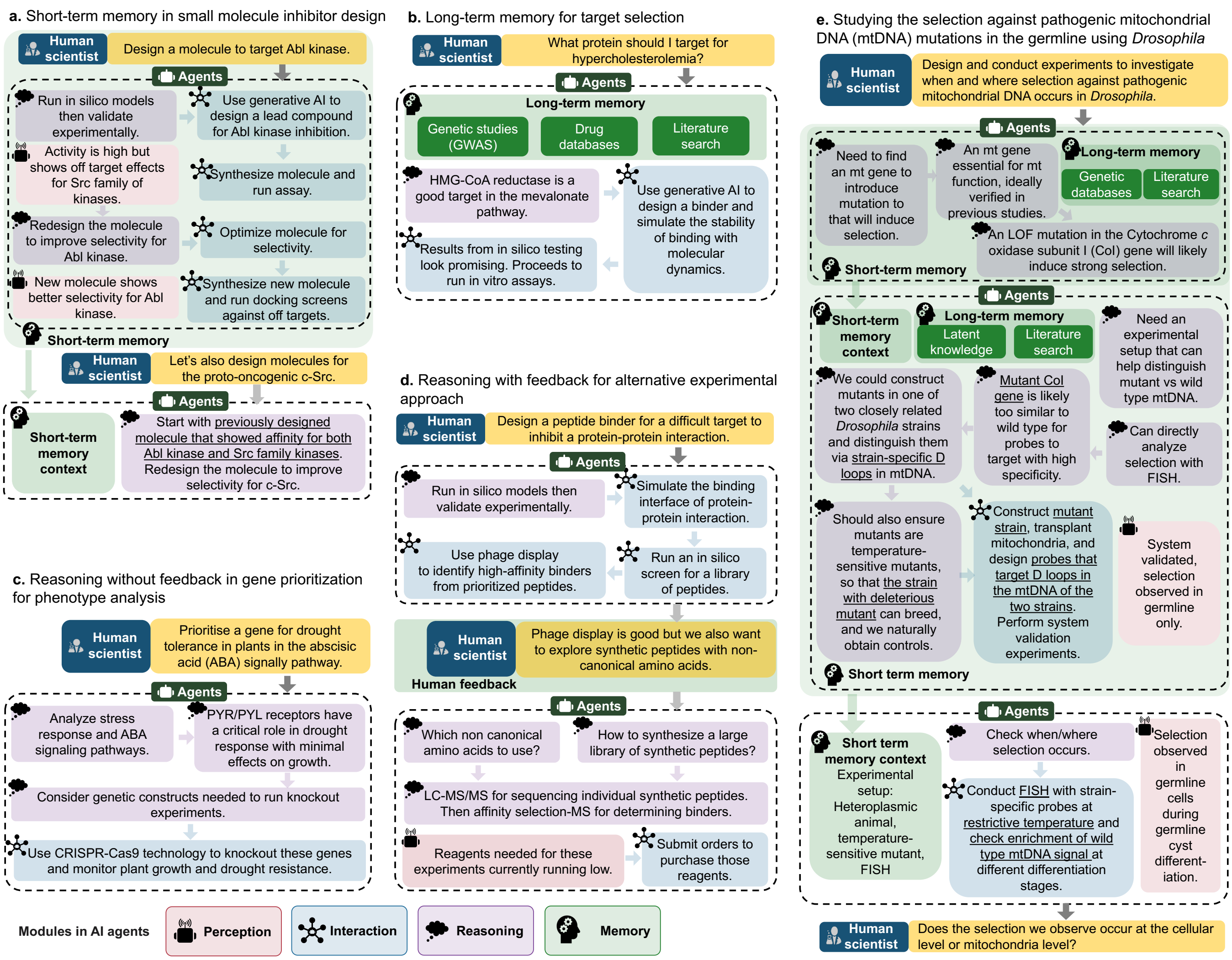

Figure 5: Components of biomedical AI agents. a. Use of a short-term memory module to recall previous relevant experiments for small molecule inhibitor design. **b.** Use of a long-term memory module to retrieve relevant information for target selection for a disease. **c.** Use of reasoning without scientist feedback in gene prioritization for phenotype analysis. **d.** Use of reasoning with feedback from scientists to select an alternative experimental approach.

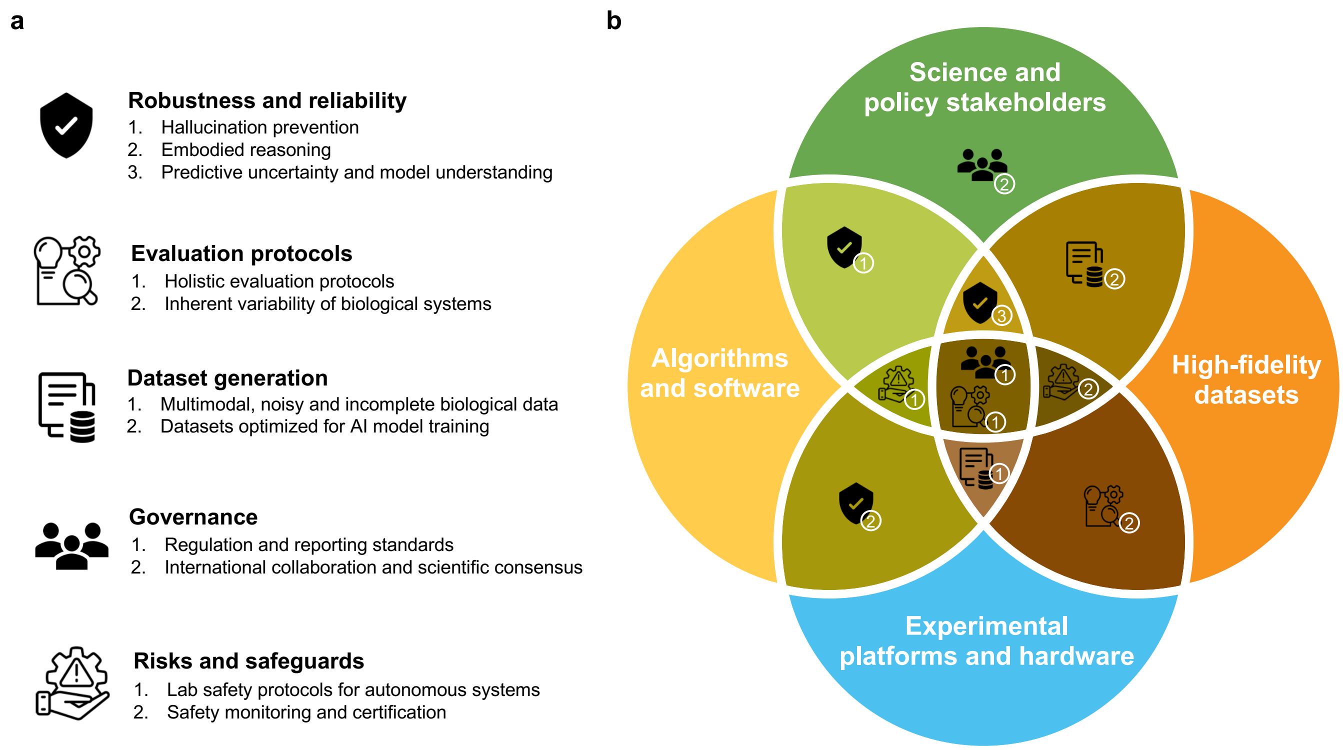

Figure 6: Challenges for AI agents in scientific discovery. a. Shown are critical challenges–including robustness and reliability, evaluation protocols, dataset generation, governance, and risks—alongside **b.** strategic approaches to address them.

Autonomy levels	Scientific discovery Hypothesis	 Experiment	 Reasoning	Scientist-AI agent roles
Level 0: **No AI agent**	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 their generation · Scientist defines the task to test hypothesis · Scientist completes tasks
Level 1: **AI agent as an assistant**	AI generates simple and narrow hypotheses that are direct composition of existing knowledge	Narrow design of experimental protocols and utilization of in silico and experimental tools	· Strong reasoning in a selected task · Multi-modal 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 agent as a collaborator**	AI 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	· Identifying pioneering discoveries · Synthesis of concepts, not a summary of findings	· Scientist proposes initial hypothesis and refines hypothesis with AI agent · AI agent defines the series of tasks to test hypothesis · AI agent completes tasks
Level 3: **AI agent as a scientist**	AI generates creative, de novo hypotheses that are indirect extrapolations from existing knowledge.	Development of experimental methods unlocking new capabilities	· Contextualizing pioneering discoveries · 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

Table 1: Levels of autonomy in AI agents. 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.

Term	Description
Multi-modal 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 of models, such as LLMs, to achieve new tasks based on the context provided within contextual prompts, without requiring explicit model training

Table 2: Glossary of key machine learning 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 in order 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
Cryo-electron 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

Table 3: Glossary of key biological terms.

References

1. Hiroaki Kitano. Nobel turing challenge: creating the engine for scientific discovery. *NPJ systems biology and applications*, 7(1):29, 2021.

2. Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

3. Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew White, and Philippe Schwaller. Augmenting large language models with chemistry tools. In *NeurIPS 2023 AI for Science Workshop*, 2023. URL `https://openreview.net/forum?id=wdGIL6lx3l`.

4. Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

5. Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

6. Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

7. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

8. Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

9. Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

10. Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. Chatgpt for robotics: Design principles and model abilities. *Microsoft Auton. Syst. Robot. Res*, 2:20, 2023.

11. Significant Gravitas. Autogpt, 2023. URL `https://agpt.co`.

12. Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=WE_vluYUL-X`.

13. Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

14. Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

15. Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530, 2023. doi: 10.1109/ICRA48891.2023.10161317.

16. Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022.

17. Mario Krenn, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, AkshatKumar Nigam, et al. On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4(12):761–769, 2022.

18. Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. Trustllm: Trustworthiness in large language models. *arXiv:2401.05561*, 2024.

19. Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. *arXiv:2309.10105*, 2023.

20. Hanzhou Li, John T Moon, Saptarshi Purkayastha, Leo Anthony Celi, Hari Trivedi, and Judy W Gichoya. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, 5(6): e333–e335, 2023.

21. Lea Goetz, Markus Trengove, Artem Trotsyuk, and Carole A Federico. Unreliable llm bioethics assistants: Ethical and pedagogical risks. *The American Journal of Bioethics*, 23(10):89–91, 2023.

22. Ashutosh Kumar, Sagarika Singh, Shiv Vignesh Murty, and Swathy Ragupathy. The ethics of interaction: Mitigating security threats in llms. *arXiv:2401.12273*, 2024.

23. Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

24. Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.

25. Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

26. Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

27. Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

28. Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, pages 1–9, 2023.

29. Michael Ku Yu, Michael Kramer, Janusz Dutkowski, Rohith Srivas, Katherine Licon, Jason F Kreisberg, Cherie T Ng, Nevan Krogan, Roded Sharan, and Trey Ideker. Translation of genotype to phenotype by a hierarchy of cell subsystems. *Cell systems*, 2(2):77–88, 2016.

30. Dinesh Singh, Phillip G Febbo, Kenneth Ross, Donald G Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A Renshaw, Anthony V D'Amico, Jerome P Richie, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002.

31. Margaret A Shipp, Ken N Ross, Pablo Tamayo, Andrew P Weng, Jeffery L Kutok, Ricardo CT Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S Pinkus, et al. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, 8(1):68–74, 2002.

32. Brent M Kuenzi, Jisoo Park, Samson H Fong, Kyle S Sanchez, John Lee, Jason F Kreisberg, Jianzhu Ma, and Trey Ideker. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer cell*, 38(5):672–684, 2020.

33. Feng Ren, Alex Aliper, Jian Chen, Heng Zhao, Sujata Rao, Christoph Kuppe, Ivan V. Ozerov, Man Zhang, Klaus Witte, Chris Kruse, Vladimir Aladinskiy, Yan Ivanenkov, Daniil Polykovskiy, Yanyun Fu, Eugene Babin, Junwen Qiao, Xing Liang, Zhenzhen Mou, Hui Wang, Frank W. Pun, Pedro Torres Ayuso, Alexander Veviorskiy, Dandan Song, Sang Liu, Bei Zhang, Vladimir Naumov, Xiaoqiang Ding, Andrey Kukharenko, Evgeny Izumchenko, and Alex Zhavoronkov. A small-molecule tnik inhibitor targets fibrosis in preclinical and clinical models. *Nature Biotechnology*, March 2024. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-024-02143-0. URL `https://www.nature.com/articles/s41587-024-02143-0`.

34. Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

35. Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

36. 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, 2012.

37. David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl_1):D668–D672, 2006.

38. Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

39. John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

40. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

41. Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

42. Anna Gaulton, Anne Hersey, Michał Nowotka, A Patricia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, et al. The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954, 2017.

43. Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, pages 1–4, 2023.

44. Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

45. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

46. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

47. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

48. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

49. Alex Hernández-García, Nikita Saxena, Moksh Jain, Cheng-Hao Liu, and Yoshua Bengio. Multi-fidelity active learning with GFlownets, 2024. URL `https://openreview.net/forum?id=3QR230r11w`.

50. Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=TG8KACxEON`.

51. Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9): 1038–1040, 2019.

52. Brian L. Hie and Kevin K. Yang. Adaptive machine learning for protein engineering. *Current Opinion in Structural Biology*, 72:145–152, February 2022. ISSN 0959440X. doi: 10.1016/j.sbi.2021.11.002. URL `https://linkinghub.elsevier.com/retrieve/pii/S0959440X21001457`.

53. Isaac D. Lutz, Shunzhi Wang, Christoffer Norn, Alexis Courbet, Andrew J. Borst, Yan Ting Zhao, Annie Dosey, Longxing Cao, Jinwei Xu, Elizabeth M. Leaf, Catherine Treichel, Patrisia Litvicov, Zhe Li, Alexander D. Goodson, Paula Rivera-Sánchez, Ana-Maria Bratovianu, Minkyung Baek, Neil P. King, Hannele Ruohola-Baker, and David Baker. Top-down design of protein architectures with reinforcement learning. *Science*, 380(6642):266–273, April 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adf6591.

54. Michael Bailey, Saeed Moayedpour, Ruijiang Li, Alejandro Corrochano-Navarro, Alexander Kötter, Lorenzo Kogler-Anele, Saleh Riahi, Christoph Grebner, Gerhard Hessler, Hans Matter, et al. Deep batch active learning for drug discovery. *bioRxiv*, pages 2023–07, 2023.

55. Ava P. Soleimany, Alexander Amini, Samuel Goldman, Daniela Rus, Sangeeta N. Bhatia, and Connor W. Coley. Evidential deep learning for guided molecular property prediction and discovery. *ACS Central Science*, 7(8):1356–1367, August 2021. ISSN 2374-7943, 2374-7951. doi: 10.1021/acscentsci.1c00546.

56. Jiaqi Zhang, Louis Cammarata, Chandler Squires, Themistoklis P. Sapsis, and Caroline Uhler. Active learning for optimal intervention design in causal models. *Nature Machine Intelligence*, 5(10):1066–1075, October 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00719-0.

57. Adam Yala, Peter G. Mikhael, Constance Lehman, Gigin Lin, Fredrik Strand, Yung-Liang Wan, Kevin Hughes, Siddharth Satuluru, Thomas Kim, Imon Banerjee, Judy Gichoya, Hari Trivedi, and Regina Barzilay. Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nature Medicine*, 28(1):136–143, January 2022. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-021-01599-w. URL `https://www.nature.com/articles/s41591-021-01599-w`.

58. Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=1i6ZCvflQJ`. Survey Certification.

59. Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):1–26, 2024.

60. Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=gEZrGCozdqR`.

61. Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

62. Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

63. Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

64. Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022.

65. Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, et al. Health system-scale language models are all-purpose prediction engines. *Nature*, pages 1–6, 2023.

66. Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

67. Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

68. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

69. Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. In *Intrinsically-Motivated and Open-Ended Learning Workshop @NeurIPS2023*, 2023. URL `https://openreview.net/forum?id=nfx5IutEed`.

70. Chrisantha Fernando, Dylan Sunil Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution, 2024. URL `https://openreview.net/forum?id=HKkiX32Zw1`.

71. Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023.

72. Yann LeCun. A path towards autonomous machine intelligence. *Open Review*, 62(1), 2022.

73. Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*, 2023.

74. Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*, 2023.

75. Lauren M Sanders, Ryan T Scott, Jason H Yang, Amina Ann Qutub, Hector Garcia Martin, Daniel C Berrios, Jaden JA Hastings, Jon Rask, Graham Mackintosh, Adrienne L Hoarfrost, et al. Biological research and self-driving labs in deep space supported by artificial intelligence. *Nature Machine Intelligence*, 5(3):208–219, 2023.

76. Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, et al. Advancing mathematics by guiding human intuition with ai. *Nature*, 600(7887):70–74, 2021.

77. David J Glass and Ned Hall. A brief history of the hypothesis. *Cell*, 134(3):378–381, 2008.

78. Yang Lim, Lukas Tamayo-Orrego, Ernst Schmid, Zygimante Tarnauskaite, Olga V Kochenova, Rhian Gruar, Sachiko Muramatsu, Luke Lynch, Aitana Verdu Schlie, Paula L Carroll, et al. In silico protein interaction screening uncovers donson's role in replication initiation. *Science*, 381(6664):eadi3448, 2023.

79. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

80. Juexiao Zhou, Bin Zhang, Xiuying Chen, Haoyang Li, Xiaopeng Xu, Siyuan Chen, and Xin Gao. Automated bioinformatics analysis via autoba. *arXiv preprint arXiv:2309.03242*, 2023.

81. Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2023.

82. Xiuyuan Hu, Guoqing Liu, Yang Zhao, and Hao Zhang. De novo drug design using reinforcement learning with multiple gpt agents. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

83. Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of agi: Operationalizing progress on the path to agi. *arXiv preprint arXiv:2311.02462*, 2023.

84. Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191, March 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00465-9. URL `https://www.nature.com/articles/s42256-022-00465-9`.

85. Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuosheng Zhang, Arman Cohan, Zhiyong Lu, and Mark Gerstein. Prioritizing safeguarding over autonomy: Risks of llm agents for science. *arXiv preprint arXiv:2402.04247*, 2024.

86. David Baker and George Church. Protein design meets biosecurity, 2024.

87. Andries T Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.*, 27(2):e1608, June 2018.

88. Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59, Aug 2021. ISSN 2662-8449. doi: 10.1038/s43586-021-00056-9. URL `https://doi.org/10.1038/s43586-021-00056-9`.

89. Felix W Frueh. Real-world clinical effectiveness, regulatory transparency and payer coverage: three ingredients for translating pharmacogenomics into clinical practice. *Pharmacogenomics*, 11(5):657–660, May 2010.

90. CP Panayiotopoulos. *The Epilepsies: Seizures, Syndromes and Management*. Bladon Medical Publishing, Oxfordshire, UK, 2005.

91. International League Against Epilepsy Consortium on Complex Epilepsies. Gwas meta-analysis of over 29,000 people with epilepsy identifies 26 risk loci and subtype-specific genetic architecture. *Nature Communications*, 9(1):1–12, 2018.

92. Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

93. Regina Gamirova, Elena Shagimardanova, Takehiro Sato, Takayuki Kannon, Rimma Gamirova, and Atsushi Tajima. Identification of potential disease-associated variants in idiopathic generalized epilepsy using targeted sequencing. *Journal of Human Genetics*, 69(2):59–67, Feb 2024. ISSN 1435-232X. doi: 10.1038/s10038-023-01208-3. URL `https://doi.org/10.1038/s10038-023-01208-3`.

94. Karen L Oliver, Ingrid E Scheffer, Mark F Bennett, Bronwyn E Grinton, Melanie Bahlo, and Samuel F Berkovic. Genes4Epilepsy: An epilepsy gene resource. *Epilepsia*, 64(5):1368–1375, May 2023.

95. Rebecca J Salowe, Roy Lee, Selam Zenebe-Gete, Marquis Vaughn, Harini V Gudiseva, Maxwell Pistilli, Ava Kikut, Emily Becker, David W Collins, Jie He, Sayaka Merriam, Kristen Mulvihill, Nora Laberee, Sara Lomax-Reese, Windell Murphy, Jeffrey Henderer, Venkata R M Chavali, Qi N Cui, Ahmara G Ross, Victoria Addis, Prithvi S Sankar, Eydie Miller-Ellis, Maureen G Maguire, and Joan M O'Brien. Recruitment strategies and lessons learned from a large genetic study of African Americans. *PLOS Glob. Public Health*, 2(8):e0000416, August 2022.

96. Brahim Aissani. Confounding by linkage disequilibrium. *Journal of Human Genetics*, 59(2):110–115, 2014. ISSN 1435-232X. doi: 10.1038/jhg.2013.130. URL `https://doi.org/10.1038/jhg.2013.130`.

97. Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander Van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. The human cell atlas. *eLife*, 6:e27041, December 2017. ISSN 2050-084X. doi: 10.7554/eLife.27041. URL `https://elifesciences.org/articles/27041`.

98. Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, David L. Lahr, Jodi E. Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C. Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M. Enache, Federica Piccioni, Sarah A. Johnson, Nicholas J. Lyons, Alice H. Berger, Alykhan F. Shamji, Angela N. Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y. Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S. Gray, Paul A. Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J. Haggarty, Lucienne V. Ronco, Jesse S. Boehm, Stuart L. Schreiber, John G. Doench, Joshua A. Bittker, David E. Root, Bang Wong, and Todd R. Golub. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, November 2017. ISSN 00928674. doi: 10.1016/j.cell.2017.10.049. URL `https://linkinghub.elsevier.com/retrieve/pii/S0092867417313090`.

99. Dylan C. Mitchell, Miljan Kuljanin, Jiaming Li, Jonathan G. Van Vranken, Nathan Bulloch, Devin K. Schweppe, Edward L. Huttlin, and Steven P. Gygi. A proteome-wide atlas of drug mechanism of action.

Nature Biotechnology, 41(6):845–857, June 2023. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-022-01539-0. URL `https://www.nature.com/articles/s41587-022-01539-0`.

100. Mahmoud Ghandi, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, Christopher C. Lo, E. Robert McDonald, Jordi Barretina, Ellen T. Gelfand, Craig M. Bielski, Haoxin Li, Kevin Hu, Alexander Y. Andreev-Drakhlin, Jaegil Kim, Julian M. Hess, Brian J. Haas, François Aguet, Barbara A. Weir, Michael V. Rothberg, Brenton R. Paolella, Michael S. Lawrence, Rehan Akbani, Yiling Lu, Hong L. Tiv, Prafulla C. Gokhale, Antoine De Weck, Ali Amin Mansour, Coyin Oh, Juliann Shih, Kevin Hadi, Yanay Rosen, Jonathan Bistline, Kavitha Venkatesan, Anupama Reddy, Dmitriy Sonkin, Manway Liu, Joseph Lehar, Joshua M. Korn, Dale A. Porter, Michael D. Jones, Javad Golji, Giordano Caponigro, Jordan E. Taylor, Caitlin M. Dunning, Amanda L. Creech, Allison C. Warren, James M. McFarland, Mahdi Zamanighomi, Audrey Kauffmann, Nicolas Stransky, Marcin Imielinski, Yosef E. Maruvka, Andrew D. Cherniack, Aviad Tsherniak, Francisca Vazquez, Jacob D. Jaffe, Andrew A. Lane, David M. Weinstock, Cory M. Johannessen, Michael P. Morrissey, Frank Stegmeier, Robert Schlegel, William C. Hahn, Gad Getz, Gordon B. Mills, Jesse S. Boehm, Todd R. Golub, Levi A. Garraway, and William R. Sellers. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757):503–508, May 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1186-3. URL `https://www.nature.com/articles/s41586-019-1186-3`.

101. Srinivas Niranj Chandrasekaran, Beth A Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Biorxiv*, pages 2022–01, 2022.

102. Marta M Fay, Oren Kraus, Mason Victors, Lakshmanan Arumugam, Kamal Vuggumudi, John Urbanik, Kyle Hansen, Safiye Celik, Nico Cernek, Ganesh Jagannathan, et al. Rxrx3: Phenomics map of biology. *bioRxiv*, pages 2023–02, 2023.

103. Irene De Teresa-Trueba, Sara K. Goetz, Alexander Mattausch, Frosina Stojanovska, Christian E. Zimmerli, Mauricio Toro-Nahuelpan, Dorothy W. C. Cheng, Fergus Tollervey, Constantin Pape, Martin Beck, Alba Diz-Muñoz, Anna Kreshuk, Julia Mahamid, and Judith B. Zaugg. Convolutional networks for supervised mining of molecular patterns within cellular context. *Nature Methods*, 20(2):284–294, February 2023. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-022-01746-2. URL `https://www.nature.com/articles/s41592-022-01746-2`.

104. Oda Helene Schiøtz, Christoph JO Kaiser, Sven Klumpe, Dustin R Morado, Matthias Poege, Jonathan Schneider, Florian Beck, David P Klebl, Christopher Thompson, and Jürgen M Plitzko. Serial lift-out: sampling the molecular anatomy of whole organisms. *Nature Methods*, pages 1–9, 2023.

105. Emma Lundberg and Georg H. H. Borner. Spatial proteomics: a powerful discovery tool for cell biology. *Nature Reviews Molecular Cell Biology*, 20(5):285–302, May 2019. ISSN 1471-0072, 1471-0080. doi: 10.1038/s41580-018-0094-y. URL `https://www.nature.com/articles/s41580-018-0094-y`.

106. Nathan H. Cho, Keith C. Cheveralls, Andreas-David Brunner, Kibeom Kim, André C. Michaelis, Preethi Raghavan, Hirofumi Kobayashi, Laura Savy, Jason Y. Li, Hera Canaj, James Y. S. Kim, Edna M. Stewart, Christian Gnann, Frank McCarthy, Joana P. Cabrera, Rachel M. Brunetti, Bryant B. Chhun, Greg Dingle, Marco Y. Hein, Bo Huang, Shalin B. Mehta, Jonathan S. Weissman, Rafael Gómez-Sjöberg, Daniel N. Itzhak, Loïc A. Royer, Matthias Mann, and Manuel D. Leonetti. Opencell: Endogenous tagging for the cartography of human cellular organization. *Science*, 375(6585):eabi6983, March 2022. ISSN 0036-8075,

1095-9203. doi: 10.1126/science.abi6983. URL
https://www.science.org/doi/10.1126/science.abi6983.

107. Graham T. Johnson, Eran Agmon, Matthew Akamatsu, Emma Lundberg, Blair Lyons, Wei Ouyang, Omar A. Quintero-Carmona, Megan Riel-Mehan, Susanne Rafelski, and Rick Horwitz. Building the next generation of virtual cells to understand cellular biology. *Biophysical Journal*, 122(18):3560–3569, September 2023. ISSN 00063495. doi: 10.1016/j.bpj.2023.04.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0006349523002369.

108. Michelle M Li, Yepeng Huang, Marissa Sumathipala, Man Qing Liang, Alberto Valdeolivas, Ashwin N Ananthakrishnan, Katherine Liao, Daniel Marbach, and Marinka Zitnik. Contextualizing protein representations using deep learning on protein networks and single-cell data. *Nature Methods*, 2024.

109. Naruki Yoshikawa, Kourosh Darvish, Mohammad Ghazi Vakili, Animesh Garg, and Alán Aspuru-Guzik. Digital pipette: open hardware for liquid transfer in self-driving laboratories. *Digital Discovery*, 2(6): 1745–1751, 2023. doi: 10.1039/D3DD00115F. URL https://pubs.rsc.org/en/content/articlelanding/2023/dd/d3dd00115f.

110. Liang Chang, Nancy Y. Jung, Adel Atari, Diego J. Rodriguez, Devishi Kesar, Tian-Yu Song, Matthew G. Rees, Melissa Ronan, Ruitong Li, Paloma Ruiz, Saireudee Chaturantabut, Takahiro Ito, Laurens M. Van Tienen, Yuen-Yi Tseng, Jennifer A. Roth, and William R. Sellers. Systematic profiling of conditional pathway activation identifies context-dependent synthetic lethalities. *Nature Genetics*, 55(10):1709–1720, October 2023. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-023-01515-7. URL https://www.nature.com/articles/s41588-023-01515-7.

111. Andrew J. C. Russell, Jackson A. Weir, Naeem M. Nadaf, Matthew Shabet, Vipin Kumar, Sandeep Kambhampati, Ruth Raichur, Giovanni J. Marrero, Sophia Liu, Karol S. Balderrama, Charles R. Vanderburg, Vignesh Shanmugam, Luyi Tian, J. Bryan Iorgulescu, Charles H. Yoon, Catherine J. Wu, Evan Z. Macosko, and Fei Chen. Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. *Nature*, 625(7993):101–109, January 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06837-4. URL https://www.nature.com/articles/s41586-023-06837-4.

112. Lotta Wik, Niklas Nordberg, John Broberg, Johan Björkesten, Erika Assarsson, Sara Henriksson, Ida Grundberg, Erik Pettersson, Christina Westerberg, Elin Liljeroth, Adam Falck, and Martin Lundberg. Proximity extension assay in combination with next-generation sequencing for high-throughput proteome-wide analysis. *Molecular and Cellular Proteomics*, 20:100168, 2021. ISSN 15359476. doi: 10.1016/j.mcpro.2021.100168. URL https://linkinghub.elsevier.com/retrieve/pii/S1535947621001407.

113. Florian A. Rosenberger, Marvin Thielert, Maximilian T. Strauss, Lisa Schweizer, Constantin Ammar, Sophia C. Mädler, Andreas Metousis, Patricia Skowronek, Maria Wahle, Katherine Madden, Janine Gote-Schniering, Anna Semenova, Herbert B. Schiller, Edwin Rodriguez, Thierry M. Nordmann, Andreas Mund, and Matthias Mann. Spatial single-cell mass spectrometry defines zonation of the hepatocyte proteome. *Nature Methods*, 20(10):1530–1536, October 2023. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-023-02007-6. URL https://www.nature.com/articles/s41592-023-02007-6.

114. Yang Liu, Marcello DiStasio, Graham Su, Hiromitsu Asashima, Archibald Enninful, Xiaoyu Qin, Yanxiang Deng, Jungmin Nam, Fu Gao, Pino Bordignon, Marco Cassano, Mary Tomayko, Mina Xu, Stephanie Halene, Joseph E. Craft, David Hafler, and Rong Fan. High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial cite-seq. *Nature Biotechnology*, 41(10):

1405–1409, October 2023. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-023-01676-0. URL `https://www.nature.com/articles/s41587-023-01676-0`.

115. Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7): 1853–1866.e17, December 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.11.038. URL `https://linkinghub.elsevier.com/retrieve/pii/S0092867416316105`.

116. Loιc Binan, Serwah Danquah, Vera Valakh, Brooke Simonton, Jon Bezney, Ralda Nehme, Brian Cleary, and Samouil L Farhi. Simultaneous crispr screening and spatial transcriptomics reveals intracellular, intercellular, and functional transcriptional circuits. *Biorxiv*, 2023.

117. Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi S. Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=kKF8_K-mBbS`.

118. Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

119. Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

120. Felix Wong, Erica J Zheng, Jacqueline A Valeri, Nina M Donghia, Melis N Anahtar, Satotaka Omori, Alicia Li, Andres Cubillos-Ruiz, Aarti Krishnan, Wengong Jin, et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, pages 1–9, 2023.

121. Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, pages 1–6, 2023.

122. Chi V Dang, E Premkumar Reddy, Kevan M Shokat, and Laura Soucek. Drugging the'undruggable'cancer targets. *Nature Reviews Cancer*, 17(8):502–508, 2017.

123. Toby Lieber, Swathi P Jeedigunta, Jonathan M Palozzi, Ruth Lehmann, and Thomas R Hurd. Mitochondrial fragmentation drives selective removal of deleterious mtdna in the germline. *Nature*, 570 (7761):380–384, 2019.

124. Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

125. Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=w0H2xGHlkw`.

126. Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553,

Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.49. URL `https://aclanthology.org/2023.emnlp-demo.49`.

127. Mackenzie Weygandt Mathis and Alexander Mathis. Deep learning tools for the measurement of animal behavior in neuroscience. *Current Opinion in Neurobiology*, 60:1–11, February 2020. ISSN 09594388. doi: 10.1016/j.conb.2019.10.008. URL `https://linkinghub.elsevier.com/retrieve/pii/S0959438819301151`.

128. Jiakun Chen, Kira E. Poskanzer, Marc R. Freeman, and Kelly R. Monk. Live-imaging of astrocyte morphogenesis and function in zebrafish neural circuits. *Nature Neuroscience*, 23(10):1297–1306, October 2020. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-020-0703-x. URL `https://www.nature.com/articles/s41593-020-0703-x`.

129. Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-e: An embodied multimodal language model. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/driess23a.html`.

130. Jiaming Li, Zhenying Cai, Laura Pontano Vaites, Ning Shen, Dylan C Mitchell, Edward L Huttlin, Joao A Paulo, Brian L Harry, and Steven P Gygi. Proteome-wide mapping of short-lived proteins in human cells. *Molecular cell*, 81(22):4722–4735, 2021.

131. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

132. Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=1tZbq88f27`.

133. Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

134. Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=vvoWPYqZJA`.

135. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

136. Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL `https://www.adept.ai/blog/fuyu-8b`.

137. Stefan Kopp and Nicole Krämer. Revisiting human-agent communication: The importance of joint co-construction and understanding mental states. *Frontiers in Psychology*, 12:580955, 2021.

138. Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and brian ichter. Inner monologue: Embodied reasoning through planning with language models. In *6th Annual Conference on Robot Learning*, 2022. URL `https://openreview.net/forum?id=3R3Pz5i0tye`.

139. Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

140. Nathalia Nascimento, Paulo Alencar, and Donald Cowan. Self-adaptive large language model (llm)-based multiagent systems. In *2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, pages 104–109. IEEE, 2023.

141. Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

142. Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=VtmBAGCN7o`.

143. Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=EnXJfQqy0K`.

144. Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

145. Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*, 2023.

146. Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models. *arXiv preprint arXiv:2307.04738*, 2023.

147. Swarnadeep Saha, Peter Hase, and Mohit Bansal. Can language models teach weaker agents? teacher explanations improve students via theory of mind. *arXiv preprint arXiv:2306.09299*, 2023.

148. Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986*, 2023.

149. Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, 2022.

150. Aaron Parisi, Yao Zhao, and Noah Fiedel. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255*, 2022.

151. Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=Yacmpz84TH`.

152. Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

153. Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving AI tasks with chatGPT and its friends in hugging face. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=yHdTscY6Ci`.

154. Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*, 2023.

155. Connor W Coley, Dale A Thomas III, Justin AM Lummiss, Jonathan N Jaworski, Christopher P Breen, Victor Schultz, Travis Hart, Joshua S Fishman, Luke Rogers, Hanyu Gao, et al. A robotic platform for flow synthesis of organic compounds informed by ai planning. *Science*, 365(6453):eaax1566, 2019.

156. Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

157. Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

158. Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

159. Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.

160. Xuanhe Zhou, Guoliang Li, and Zhiyuan Liu. Llm as dba. *arXiv preprint arXiv:2308.05481*, 2023.

161. Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. Ghost in the minecraft: Generally capable agents for open-world enviroments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023.

162. Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022.

163. Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (17):19724–19731, Mar. 2024. doi: 10.1609/aaai.v38i17.29946. URL `https://ojs.aaai.org/index.php/AAAI/article/view/29946`.

164. Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=OUIFPHEgJU`.

165. Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022.

166. Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024.

167. Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In *7th Annual Conference on Robot Learning*, 2023. URL `https://openreview.net/forum?id=wMpOMO0Ss7a`.

168. Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

169. Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.112. URL `https://aclanthology.org/2023.findings-acl.112`.

170. Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

171. Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023.

172. Gautier Dagan, Frank Keller, and Alex Lascarides. Dynamic planning with a llm. *arXiv preprint arXiv:2308.06391*, 2023.

173. Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark Gerstein, Rui Wang, Gongshen Liu, et al. Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. *arXiv preprint arXiv:2311.11797*, 2023.

174. Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2024.

175. Shreyas Sundara Raman, Vanya Cohen, Eric Rosen, Ifrah Idrees, David Paulius, and Stefanie Tellex. Planning with large language models via corrective re-prompting. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

176. Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=5Xc1ecxO1h`.

177. Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. Recmind: Large language model powered agent for recommendation. *arXiv preprint arXiv:2308.14296*, 2023.

178. Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=WZH7099tgfM`.

179. Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=1PL1NIMMrw`.

180. Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690, 2024.

181. Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL `https://openreview.net/forum?id=VTWWvYtF1R`.

182. Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=S37hOerQLB`.

183. Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023.

184. Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=KuPixIqPiq`.

185. R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*, 2023.

186. Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*, 2023.

187. Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

188. Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. *arXiv preprint arXiv:2308.16175*, 2023.

189. Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 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*, 2023. URL `https://openreview.net/forum?id=g3faCfrwm7`.

190. Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=VD-AYtP0dve`.

191. Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

192. Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. URL `http://jmlr.org/papers/v9/shafer08a.html`.

193. Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL `https://aclanthology.org/2023.findings-acl.847`.

194. Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=hTEGyKf0dZ`.

195. Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=jA235JGM09`.

196. Rishi Bommasani, Percy Liang, and Tony Lee. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023. doi: https://doi.org/10.1111/nyas.15007. URL `https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.15007`.

197. Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=fibxvahvs3`.

198. Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

199. Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Benchmarking large language models as ai research agents. *ArXiv*, abs/2310.03302, 2023. URL https://api.semanticscholar.org/CorpusID:263671541.

200. Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zAdUB0aCTQ.

201. Marinka Zitnik, Francis Nguyen, Bo Wang, Jure Leskovec, Anna Goldenberg, and Michael M. Hoffman. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71–91, 2019. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2018.09.012. URL https://www.sciencedirect.com/science/article/pii/S1566253518304482.

202. Office of Science and Technology Policy. Ai bill of rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/, 2023. Accessed: 02/2024.

203. Neel Guha, Christie Lawrence, Lindsey A Gailmard, Kit Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, et al. Ai regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *George Washington Law Review*, 11 2023. Forthcoming, Available at SSRN: https://ssrn.com/abstract=4634443.

204. Ziyi Yang, Shreyas S Raman, Ankit Shah, and Stefanie Tellex. Plug in the safety chip: Enforcing constraints for llm-driven robot agents. In *International Conference on Robotics and Automation*, 2024.

205. Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv:2107.03374*, 2021.

206. Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

207. Sivan Schwartz, Avi Yaeli, and Segev Shlomov. Enhancing trust in llm-based ai automation agents: New considerations and future challenges. In *International Joint Conference on Artificial Intelligence*, 2023.