The Dawn of the AI Co-Scientist: A Comprehensive Survey of Biomedical and Chemical Multimodal Agents and Their Applications

Qingyun Wang [†◇] **Yao Ge**[†] **Qingqing Zhu**[†] **Zhiyong Lu**[†]
[†]National Library of Medicine, National Institutes of Health, [◇]William & Mary
qwang16@wm.edu, {yao.ge,qingqing.zhu,zhiyong.lu}@nih.gov

Abstract

Autonomous and multimodal AI agents have emerged as transformative tools in biomedical and biochemical research. These AI co-scientists are capable of autonomously performing a range of complex tasks in the scientific research lifecycle, while interacting with diverse modalities (e.g., molecular structures, biomedical images, omics sequences, and structured clinical records). This survey presents the first systematic review of over 80 papers on multimodal agentic AI in biomedical and chemical research. We propose a taxonomy of biomedical and chemical multimodal agents, covering model modalities, multimodal agent learning, multimodal agent inference, existing applications, and the current landscape of benchmarks and metrics. We provide a detailed review of representative approaches and datasets in each category. This survey serves as a valuable resource for researchers interested in interdisciplinary collaboration within biomedical and chemical agentic workflows.

1 Introduction

Agentic AIs in the biomedical and chemical domain are autonomous systems powered by machine learning, particularly large language models (LLMs), to perform complex tasks in research (Wang et al., 2023a; Gao et al., 2024; Huang et al., 2025b; Ghareeb et al., 2025). These agents can understand, design, and manipulate biological and chemical molecules and systems (M. Bran et al., 2024; Ghafarollahi and Buehler, 2024; Visan and Negut, 2024) by interacting with their environment (Tom et al., 2024), reducing information overload (Landhuis, 2016). Compared to traditional AI, agentic systems integrate sophisticated reasoning capabilities and can interact dynamically with multimodal datasets, including textual data, molecular structures, biomedical images, omics sequences, and structured clinical records, thereby greatly enhancing research efficiency and reducing human workload (Ghafarollahi and Buehler, 2024; Ghareeb et al., 2025; M. Bran et al., 2024).

In biomedical and chemical domains, these AI agents face unique challenges due to the complexity and scarcity of high-quality, domain-specific datasets. For instance, the long-tail distributions and the specialized jargon present difficulties for general-purpose LLMs, requiring agents that can handle precise domain knowledge and multimodal data integration (Barnett and Doubleday, 2020; Wang et al., 2024c; Lucy et al., 2023). Additionally, the complex biomedical and chemical contexts during agent inference time require models to integrate multifaceted information (Dehghani and Levin, 2024). Recent developments emphasize specialized multimodal agentic AI that simultaneously processes and integrates heterogeneous data types, significantly outperforming their unimodal counterparts (Kim et al., 2024; Huang et al., 2025a; Schouten et al., 2025). In particular, biomedical AI agents are frequently tasked with complex workflows such as automated clinical decision support (Tang et al., 2024; Li et al., 2024a), precision diagnostic reasoning (Fan et al., 2025b), bioinformatics analyses of multi-omics data (Zhou et al., 2024a; Huang et al., 2025a), and clinical trial reasoning (Yue et al., 2024). Chemical AI agents, on the other hand, predominantly focus on drug discovery (M. Bran et al., 2024; Ivanenkov et al., 2023), molecular synthesis planning (Ma, 2025), protein engineering and optimization (Ghafarollahi and Buehler, 2024; Liu et al., 2025), and materials discovery (Ghafarollahi and Buehler, 2024). These applications require agents to interpret experimental outcomes, design scientific hypotheses, and execute complex experimental plans.

While existing surveys have extensively covered general-purpose LLM-based agents or applications in broader scientific contexts (Durante et al., 2024; Luo et al., 2025; Zheng et al., 2025), a dedicated and comprehensive review focusing specifically

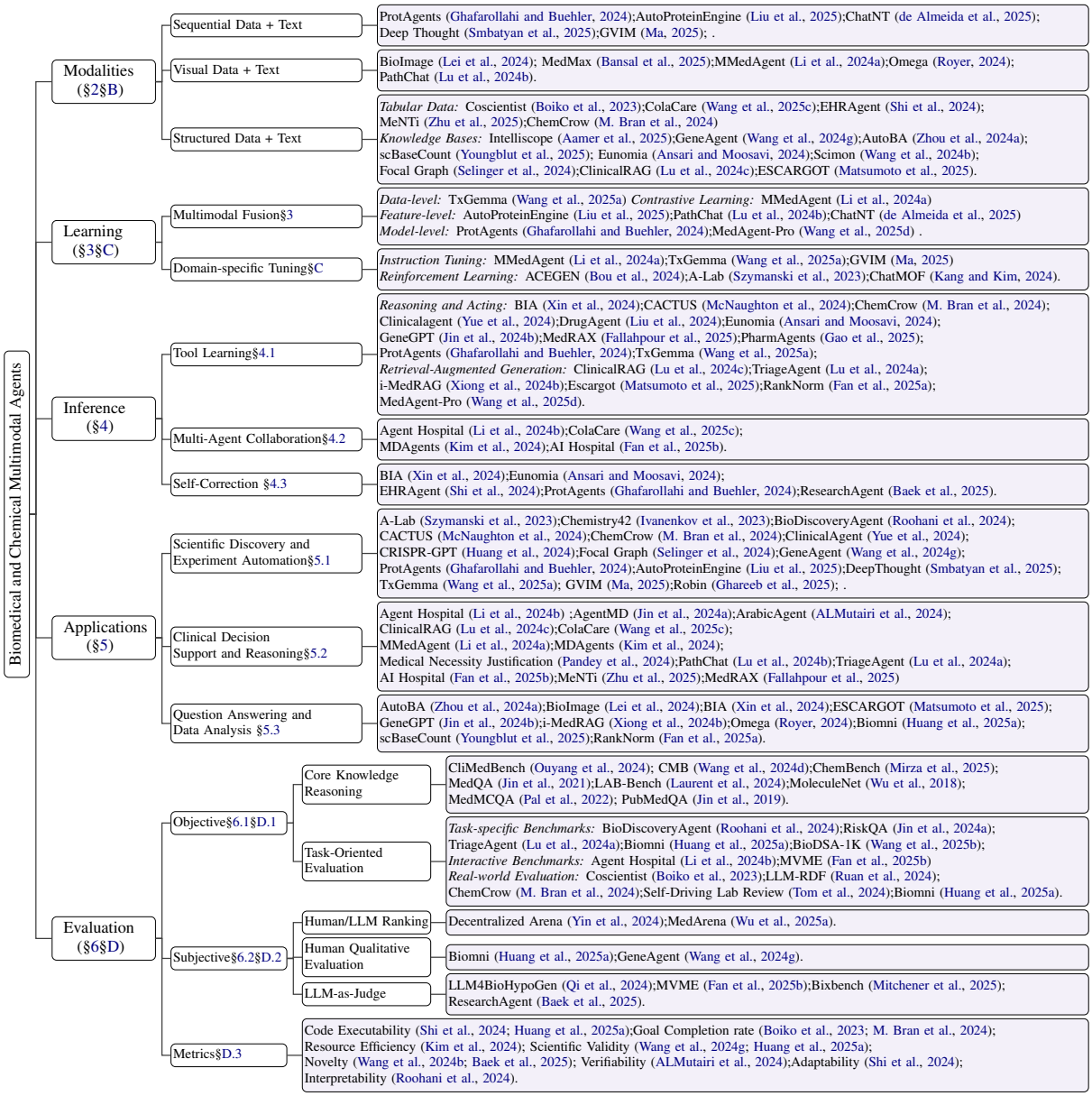

Figure 1: Taxonomy of Biomedical and chemical Multimodal Agents.

on multimodal biomedical and chemical agents remains absent. Current surveys often overlook the critical aspects unique to biomedical and chemical research (Ren et al., 2025; Gridach et al., 2025; Schouten et al., 2025), such as multimodal data integration, stringent experimental validation requirements, and domain-specific evaluation methodologies. Gao et al. (2024) presents a vision of "AI scientists" as collaborative partners in the biomedical domain, but their focus remains on vision rather than concrete implementation or applications. Ramos et al. (2025) discusses LLM-driven agents with a scope limited to chemistry and materials science. To bridge this gap, this paper systematically surveys over 80 papers on multimodal AI agents in the biomedical and chemical domains. Figure 1 shows the taxonomy of multimodal AI agents in the biomedical and chemical domains.

2 Preliminaries

2.1 Why AI agents in both Biomedicine and Chemistry?

We cover the progress of AI agents across both biomedicine and chemistry because of the deep connections between these fields, especially in drug discovery, where chemical innovations directly lead to biomedical breakthroughs. In chemistry, the current spotlight is on autonomous, AI-driven laboratories (self-driving labs) that design, plan, execute, and analyze experiments with min-

imal human intervention (Szymanski et al., 2023; Boiko et al., 2023; M. Bran et al., 2024). These systems integrate LLMs with robotic platforms to handle everything from synthesis planning to experiment optimization. Despite their chemistry orientation, these AI agents are designed to work across shared biological and chemical modalities, including molecules, proteins, and genes, making them adaptable for biomedical tasks like protein engineering (de Almeida et al., 2025; Liu et al., 2025), genetic analysis (Huang et al., 2024, 2025a), and therapeutic design (Shi et al., 2024; Wang et al., 2025a). For instance, a molecular property predictor developed for reaction optimization can be repurposed for predicting drug-target interactions in biomedicine (Kang and Kim, 2024). Meanwhile, in biomedicine, there is a growing trend toward autonomous biomedical discovery. including generating hypotheses (Qi et al., 2024; Baek et al., 2025), designing and running in silico experiments (Huang et al., 2025a), and even executing laboratory assays for biomarker identification and drug screening (Swanson et al., 2024). Because these agents share core components, such as molecule understanding, protein sequence interpretation, experimental planning, and robotic execution, the transition between chemistry and biomedicine is natural and seamless. We provide an overview of representative agentic systems in chemistry in Table E.8, and a comprehensive summary of biomedical agentic systems in Tables E.9–E.11.

2.2 Modalities Covered by Agents in Biomedical and Chemical Research (Table E.1)

Early agentic AI efforts focus on utilizing a single modality to process and analyze data in the biomedical and chemical domain (Roohani et al., 2024; ALMutairi et al., 2024; Li et al., 2024b; Pandey et al., 2024; Fan et al., 2025b). However, in the real world, scientists usually process data from diverse sources in their research (Acosta et al., 2022), since unimodal agents often face reporting bias (Gordon and Van Durme, 2013). Recently, researchers have been shifting towards multimodal agentic AI, which can process and integrate data from multiple data types simultaneously (Kim et al., 2024; Huang et al., 2025a), such as text, images, molecular structures, etc. Compared to their unimodal counterparts, multimodal models perform much better (Schouten et al., 2025). Multimodal foundation models (Moor et al., 2023; Pei et al., 2024; Zhang et al., 2024a,b; Hu and Quon, 2024; Luo et al., 2025) form the backbone of multimodal biomedical agents (Li et al., 2024a; Visan and Negut, 2024; Ghafarollahi and Buehler, 2024) for integrating domain knowledge across various modalities. These modalities include textual descriptions (Fan et al., 2025a), tabular data (Shi et al., 2024; Zhou et al., 2024a; Wang et al., 2025c), chemical small molecules (M. Bran et al., 2024; Ma, 2025; Smbatyan et al., 2025), DNA/RNA nucleotide sequence (de Almeida et al., 2025; Wang et al., 2025a), visual information (Lu et al., 2024b; Lei et al., 2024; Li et al., 2024a; Royer, 2024), protein amino acid sequences (Ghafarollahi and Buehler, 2024; Liu et al., 2025), and knowledge graphs (Wang et al., 2024g; Selinger et al., 2024; Aamer et al., 2025; Lu et al., 2024c; Matsumoto et al., 2025).

3 Multimodal Agentic AI Learning (Table E.2)

Multimodal integration is the foundational step for utilizing data from different modalities in Agentic AIs, as it aims to create meaningful correspondences and relationships between information derived from various modalities.

Data-level Integration. Due to the recent dramatic advances in LLMs, a straightforward method is to convert other modalities into text-based representations. For example, sequential biological data are commonly formatted using FASTA (Lipman and Pearson, 1985), which encodes DNA and RNA as nucleotide sequences and proteins as amino acid chains. Additionally, researchers often simplify small molecules as linear sequences, such as SMILES (Weininger, 1988), SELFIES (Krenn et al., 2020). A notable application of this approach is TxGemma (Wang et al., 2025a), the backbone LLM of Agentic-Tx. TxGemma is instruction-finetuned on a comprehensive dataset, the Therapeutic Data Commons (TDC) (Huang et al., 2021, 2022), which includes small molecules, proteins, nucleic acids, diseases, and cell lines. However, this strategy requires extensive multimodal corpora for pretraining or finetuning, which in turn needs significant computational and data resources.

Joint Embedding Alignment via Contrastive Learning. Many multimodal agents rely on contrastive learning, which brings representations of positive pairs of multimodal data closer together in the embedding space, while simultaneously

pushing representations of negative pairs farther apart (Chen et al., 2020). Notably, CLIP (Radford et al., 2021) architecture is widely used to align embeddings across different modalities. For example, MMedAgent (Li et al., 2024a) relies on BiomedCLIP (Zhang et al., 2025) for biomedical image classification by aligning biomedical images with corresponding textual descriptions. Similarly, Pharmagents (Gao et al., 2025) utilizes DrugCLIP (Gao et al., 2023) to align representations of binding proteins and molecules.

Feature-level Integration. To achieve better alignment across modalities, multimodal agents typically extract features from each modality independently before integrating them into a unified multimodal representation (Guarrasi et al., 2025). A common approach involves projecting features from a specialized encoder into the textual embedding space. For example, AutoProteinEngine (Liu et al., 2025) combines sequence and graph representations of proteins with a late fusion strategy by applying different encoders. Similarly, PathChat (Lu et al., 2024b) utilizes LLaVA-Med (Li et al., 2023) to project image features, encoded by a vision encoder, into the language embedding space. ChatNT (de Almeida et al., 2025) adopts a separate DNA encoder to project DNA representations into the textual space.

Model-level Integration. Apart from the above methods, multimodal AI agents improve accuracy by integrating the outputs from multiple domain-specific agents. A backbone LLM agent, such as GPT-4 (OpenAI, 2023), often functions as the "brain" or controller, which serves as a bridge between users and expertise-specialized agents. This backbone agent can interpret natural language instructions, perform reasoning, plan action sequences, interface with diverse external tools, and synthesize their outputs into coherent results. For instance, ProtAgents (Ghafarollahi and Buehler, 2024) is a multi-agent framework for protein design and analysis that employs GPT-4 as a coordinator to manage a suite of tools spanning various disciplines, including physics simulators and protein folding models. Similarly, using GPT-4 as a coordinator, MedAgent-Pro (Wang et al., 2025d) integrates specialized components, such as segmentation models (e.g., Medical SAM Adapter (Wu et al., 2025b)), grounding models (e.g., Maira-2 (Bannur et al., 2024)), and LLM-based coding tools (e.g., Copilot (GitHub, 2021)), to perform comprehensive quantitative analyses.

4 Multimodal Agentic AI Inference (Table E.3)

Prompt Enginnering (Radford et al., 2019), such as Chain-of-Thought (Wei et al., 2022), In-context learning (Brown et al., 2020), and ReAct (Yao et al., 2023), leverage the inherent capabilities of large pretrained foundation models without requiring direct modifications to model weights. They rely on carefully crafted input prompts, including task descriptions, examples, and specific instructions to guide the model's reasoning and output generation.

4.1 Tool Learning

Despite the impressive performance of LLMs in general tasks, they face challenges when applied to biomedical and chemical domains, such as predicting drug-target interactions and planning retrosynthesis. To bridge this gap, AI agents utilize external tools to enhance their capabilities. Such external tools can be categorized into two categories (Ren et al., 2025): (1) Application Programming Interfaces (APIs) and code libraries, and (2) simulators or physical platforms. APIs and code libraries are typically developed by domain experts (Jin et al., 2024b; Yue et al., 2024; Fallahpour et al., 2025). For instance, GeneGPT (Jin et al., 2024b) and TRIAGEAGENT(Lu et al., 2024a) utilize NCBI Web APIs to access biomedical databases. In chemistry, CACTUS (McNaughton et al., 2024) uses RDKit (Bento et al., 2020) to perform molecule property prediction. Recently, code-generating LLMs have also been used to create novel tools tailored to specific problems. For example, AgentMD (Jin et al., 2024a) and MeNTi (Zhu et al., 2025) generates clinical calculators from GPT-4 (OpenAI, 2023) for medical risk prediction. Simulators and physical environments offer platforms for AI agents to validate and test their outputs. For example, AtomAgents (Ghafarollahi and Buehler, 2025) analyzes molecule structures by physics simulations. In the chemical domain, Coscientist (Boiko et al., 2023) leverages GPT-4 with capabilities, including information retrieval, code execution, and robotic experiment automation. Similarly, ChemCrow (M. Bran et al., 2024) integrates 18 expert-designed tools with GPT-4 (OpenAI, 2023) as the LLM, enabling interaction with IBM RoboRXN (Pyzer-Knapp et al., 2022) in the physical world to discover and synthesize novel chromophores.

Reasoning and Acting. A widely adopted prompt-

ing framework is ReAct (Yao et al., 2023), which guides LLMs to interleave reasoning with predefined actions that interact with the external environment. The downstream applications ranging from protein discovery (Ghafarollahi and Buehler, 2024), drug discovery (McNaughton et al., 2024; Liu et al., 2024), material discovery (M. Bran et al., 2024; Ansari and Moosavi, 2024; Gao et al., 2025), gene analysis (Jin et al., 2024b), clinical trial analysis (Yue et al., 2024; Fallahpour et al., 2025), therapeutic development (Wang et al., 2025a), and bioinformatic analysis (Xin et al., 2024). Besides ReAct, recent AI agents also use code as a universal interface to interact with tools (Wang et al., 2024e). Biomni (Huang et al., 2025a) and EHRAgent (Shi et al., 2024) leverage the LLM's ability to generate code as the mechanism to interact with tools or execute complex, multi-step actions, because code enables complex logic, sequential operations, conditional execution, and interaction with diverse software environments and data. Given that LLMs are increasingly proficient at code generation, this approach provides a flexible and extensible means of granting agents new capabilities (Jin et al., 2024b).

Retrieval-Augmented Generation. Retrieval-augmented generation (RAG) (Lewis et al., 2020) represents a special case of tool learning by employing search engine for LLMs (Qu et al., 2025), where agents retrieve relevant information from external knowledge bases before generating a response. Since the generated response can be grounded in domain-specific or up-to-date data, it mitigates issues like hallucination and knowledge cutoffs (Agrawal et al., 2024). RAG is widely used in medical domain. ClinicalRAG (Lu et al., 2024c) establishes the foundation as a medical multi-agent pipeline that incorporates heterogeneous structured and unstructured medical knowledge into LLMs to reduce hallucination and improve diagnostic accuracy. Building on that, MedAgent-Pro (Wang et al., 2025d) advances this paradigm by including an RAG agent that retrieves up-to-date medical guidelines from MedlinePlus (Miller et al., 2000) during diagnostic plan generation. Escargot (Matsumoto et al., 2025) utilizes RAG pipelines to better support clinical decision-making and biomedical research by grounding model outputs in Alzheimer's knowledge graph (Romano et al., 2024). Similarly, TriageAgent (Lu et al., 2024a) uses RAG to fetch precise clinical criteria from the Emergency Severity Index handbook to support triage decision-making in the emergency department. Fan et al. (2025a) leverages RAG to expand the short medical texts into knowledge cards containing enhanced descriptive information and medical knowledge. Finally, i-MedRAG (Xiong et al., 2024b) extends standard RAG by enabling LLMs to ask iterative follow-up questions during the process.

4.2 Multi-Agent Collaboration

Interactions between multiple AI agents can lead to improved performance or collective learning (Wang et al., 2024f). For example, AI Hospital (Fan et al., 2025b) proposes a LLM-powered multi-agent framework that simulates medical interactions through collaborative diagnosis with dispute resolution to enhance diagnostic accuracy across multiple discussion iterations. CoLaCare (Wang et al., 2025c) presents a multi-agent framework to integrates LLM with domain-specific expert models to collaboratively analyze structured EHR data and bridge it with text-based reasoning for improved medical record modeling. Similarly, MDAgents (Kim et al., 2024) utilizes an adaptive decision-making framework through dynamic collaboration among AI agents based on the complexity of the medical task. Beyond that, Agent Hospital (Li et al., 2024b) constructs a virtual hospital environment by simulating patients, nurses, and doctors as LLM-powered agents. Within this simulacrum, doctor agents progressively evolve by treating vast numbers of simulated patient cases which are designed and documented by the system itself, thereby refining their diagnostic and treatment capabilities over time.

4.3 Self-Correction and Iterative Refinement

Self-correction and iterative refinement (Madaan et al., 2023; Kamoi et al., 2024) refers to the process by which an LLM refines its own output during inference to improve accuracy iteratively. ProtAgents (Ghafarollahi and Buehler, 2024) uses "Critic" agent to identify mistakes in plan proposals or execution and suggest fixes for errors. Similarly, ResearchAgent (Baek et al., 2025) proposes "ReviewingAgents" to iteratively refine research ideas based on feedback from collaborative LLM-powered reviewing agents. EHRAgent (Shi et al., 2024) and BioInformatics Agent (BIA) (Xin et al., 2024) learn from error messages and iteratively improves originally generated code by integrating feedback. Eunomia (Ansari and Moosavi, 2024) employs an iterative chain-of-verification process, systematically reviewing and correcting its own

reasoning to ensure each step logically supports accurate predictions of water stability in materials.

5 Applications (Table E.5)

With the recent breakthroughs in LLMs, AI agents tackle a wide range of tasks in research and healthcare. By integrating conversation ability with multimodal interfaces, they automate workflows that traditionally require substantial human expertise.

5.1 Scientific Discovery and Experiment Automation

AI agents are increasingly engaged in critical scientific tasks, such as hypothesis generation, therapeutic target identification, molecular and protein design, gene editing setups, and bioinformatics analyses. For example, Robin (Ghareeb et al., 2025) identifies novel therapeutic candidates such as ROCK inhibitors for dry age-related macular degeneration (AMD). In parallel, ClinicalAgent (Yue et al., 2024) focuses on clinical trial reasoning and drug development by simulating multi-agent collaboration for outcome prediction and failure analysis. In gene engineering, BioDiscoveryAgent (Roohani et al., 2024) applies genetic perturbation experiments to uncover new gene targets, while GeneAgent (Wang et al., 2024g) performs gene set knowledge discovery for advancing human functional genomics. CRISPR-GPT (Huang et al., 2024) simplifies the experimental setup for gene editing. Protein engineering is also being revolutionized. AutoProteinEngine (Liu et al., 2025) and ProtAgents (Ghafarollahi and Buehler, 2024) streamline protein engineering, covering tasks from mutation prediction to property optimization.

In chemistry, ChemCrow (M. Bran et al., 2024) and GVIM (Ma, 2025) support molecular synthesis and retrosynthesis planning. DeepThought (Smbatyan et al., 2025), TxGemma (Wang et al., 2025a), and Chemistry42 (Ivanenkov et al., 2023) advance drug discovery through virtual screening and molecular design. Selinger et al. (2024) supports autonomous drug target discovery through biomedical knowledge graphs. A-Lab (Szymanski et al., 2023) accelerates materials discovery by bridging computational predictions and laboratory synthesis. CACTUS (McNaughton et al., 2024) supports molecular property prediction and drug design by connecting LLMs with cheminformatics tools. In summary, these agents illustrate AI's growing role in enhancing efficiency and innovation in scientific research, showing how AI agents effectively streamline complex scientific tasks with minimal human intervention.

5.2 Clinical Decision Support and Reasoning

Clinical AI agents are evolving to help healthcare professionals interpret patient data and enhance diagnostic and therapeutic decision-making. Systems such as MMedAgent (Li et al., 2024a) and ClinicalRAG (Lu et al., 2024c) combine medical images, structured EHR data, and retrieval-augmented generation to improve accuracy and transparency in diverse diagnostic tasks. Meanwhile, MeNTi (Zhu et al., 2025) enables quantitative clinical assessment by operating medical calculators through nested tool calls. Several platforms go further by simulating full clinical workflows. To expand agent competence safely, Agent Hospital (Li et al., 2024b) and AI Hospital (Fan et al., 2025b) simulate virtual clinical ecosystems where multi-role agents interact with patients and collaborate on complex diagnoses through iterative dialogues and consensus reports. In specialized applications, PathChat (Lu et al., 2024b) supports slide interpretation and pathology consultation; CoLaCare (Wang et al., 2025c) taps into guidelines and structured knowledge to advise on oncology staging trial eligibility, and treatment recommendations; AgentMD (Jin et al., 2024a) automates ICU risk stratification directly from clinical notes using curated calculators. For critical workflows, TriageAgent (Lu et al., 2024a) improves emergency triage by combining LLM-driven reasoning with clinical handbook retrieval and public triage benchmarks, while MedRAX (Fallahpour et al., 2025) integrates chest X-ray analysis with multimodal LLMs to handle complex diagnostic queries.

To make medical AI more accessible, agents are now being taught to follow step-by-step clinical instructions and speak multiple languages. For low-resource contexts, ArabicAgent (ALMutairi et al., 2024) generates culturally grounded Najdi dialect medical dialogues to support doctor-patient interactions. Several multi-agent frameworks, including MDAgents (Kim et al., 2024) and Medical Necessity Justification (Pandey et al., 2024), tackle collaborative clinical decision-making, administrative justification, and argument-based reasoning by simulating multi-role discussions. Additionally, MedMax (Bansal et al., 2025) expands instruction-following capabilities to multimodal biomedical tasks for domains like radiology and pathology.

5.3 Biomedical and Chemical Question Answering and Data Analysis

Terminology normalization and biomedical knowledge grounding are the foundation of biomedical data analysis. For example, RankNorm (Fan et al., 2025a) proposes a training-free multi-agent framework that normalizes informal health mentions from social media into standard biomedical terms. Once terms are normalized, Question Answering (QA) agents can answer biological and clinical questions accurately, safely, and with clear justification. Unlike general-purpose QA systems, these agents ground their responses in biomedical knowledge, integrate tool usage, and often explain their reasoning in high-stakes scenarios. For example, i-MedRAG (Xiong et al., 2024b) enhances medical QA by enabling LLMs to iteratively generate follow-up queries, supporting multi-hop retrieval over clinical knowledge sources. Additionally, powered by NCBI Web APIs, GeneGPT (Jin et al., 2024b) allows precise retrieval of gene, variant, and sequence information for genomics tasks.

To support more complex workflows in biomedical research, recent systems integrate LLM agents with external tools to support multimodal data analysis. For example, BioImage (Lei et al., 2024) and Omega (Royer, 2024) help users with natural language-driven bioimage analysis, covering tasks such as processing, segmentation, and interactive visualization. ESCARGOT (Matsumoto et al., 2025) enhances biomedical reasoning and research design by integrating knowledge graphs and multi-agent collaboration. In bioinformatics, AutoBA (Zhou et al., 2024a) and Biomni (Huang et al., 2025a) automate complex multi-omic analysis, variant annotation, and phenotypic profiling. Tools like scBaseCount (Youngblut et al., 2025) offer vast, curated single-cell RNA-seq repositories, while BIA (BioInformatics Agent) (Xin et al., 2024) can execute full pipelines, automating tasks like single-cell RNA-seq data processing and reporting. These agents highlight the potential of agent-tool integration to accelerate biomedical research and analysis by enabling end-to-end automation through conversational interfaces.

6 Evaluations and Benchmark

6.1 Objective Evaluation (Table E.6)

Core Knowledge Reasoning. The first step is to evaluate the core knowledge and reasoning capabilities of AI agents.

In the clinical domain, a common method is to test AI systems against the same standards used for human physicians. For instance, MedQA (Jin et al., 2021) is based on the US medical licensing exams, MedMCQA (Pal et al., 2022) draws from Indian medical entrance exams, while CliMedBench (Ouyang et al., 2024) and CMB (Wang et al., 2024d) are based on Chinese medical licensing exams and clinical cases. Beyond those, QA datasets also focus on an agent's ability to retrieve and synthesize information from scientific literature, databases, and its parametric knowledge. For instance, PubMedQA (Jin et al., 2019) evaluates reading comprehension of PubMed abstracts, GeneHop (Jin et al., 2024b) tests multi-hop reasoning via API calls, and LAB-Bench (Laurent et al., 2024) challenges models to interact with structured biological databases and reason over DNA and protein sequences. Recently, ChemBench (Mirza et al., 2025) curated more than 2,700 question–answer pairs from academic sources to evaluate the chemical knowledge and reasoning abilities of state-of-the-art LLMs against the expertise of chemists. In addition to QA tasks, models are also evaluated using classification and regression benchmarks, such as MoleculeNet (Wu et al., 2018), which includes over 700,000 compounds assessed across various properties for both classification and regression tasks. While strong performance on these benchmarks is essential, it does not guarantee proficiency in complex, real-world tasks (Fan et al., 2025b).

Task-Oriented Evaluation. While earlier benchmarks focus on what an AI agent *knows*, its real-world utility depends on what it can *do*. Therefore, researchers have moved beyond static knowledge tests to evaluate agents on complex, multi-step tasks that mimic real scientific and clinical workflows. These evaluations follow a clear progression of difficulty: the evaluation of task completion, the performance within interactive, simulated environments, and real-world outcomes. Task-specific benchmarks aim to assess an agent's ability to perform a complete, end-to-end task (Roohani et al., 2024; Huang et al., 2025a; Wang et al., 2025b; Lu et al., 2024a). The interactive benchmarks assess not just the final outcome but the entire process of interaction, decision-making, and adaptation for AI agents (Li et al., 2024b; Fan et al., 2025b). Finally, real-world outcomes, such as wet-lab experiments, represent the ultimate benchmark for testing AI agents (Huang et al., 2025a; Tom et al., 2024; Boiko et al., 2023; Ruan et al., 2024; M. Bran et al.,

2024). They evaluate agents that interact with physical laboratory hardware, grounding evaluation in real-world outcomes by designing a Self-Driving Laboratory (SDL).

6.2 Subjective Evaluation (Table E.7)

Because AI agent applications are highly complex, there can be cases where no evaluation datasets exist, or where obtaining quantitative evaluation metrics proves difficult (Wang et al., 2024a). Therefore, researchers often depend on human judgment or employ LLMs as judges to evaluate agent effectiveness. Researchers use Human or LLMs to test and compare top-performing LLMs on their performance, such as MedArena (Wu et al., 2025a) and Decentralized Arena (Yin et al., 2024). Additionally, for many complex agentic tasks, where purely quantitative metrics are insufficient or nonexistent, their evaluation relies on direct, structured judgment by human experts (Ethayarajh and Jurafsky, 2022; Wang et al., 2024g; M. Bran et al., 2024). To scale up the evaluation, researchers also rely on LLM-as-a-Judge to assess the quality of AI-generated responses (Zheng et al., 2023; Qi et al., 2024; Mitchener et al., 2025).

7 Challenges and Future Directions

While AI agents indicate a new era in biomedical and chemical research, their deployment faces critical challenges in data quality, reliability, reasoning, evaluation, and safety governance.

Data Quality and Scarcity. Training multimodal AI systems requires large, well-annotated datasets pairing diverse modalities. However, data scarcity remains pervasive due to experimental variability, fragmented documentation, and limited dataset sizes. For example, over 20% of Therapeutics Data Commons datasets contain fewer than 1,000 data points (Huang et al., 2021, 2022). Data biases and errors further complicate reliability. For instance, Walters (2023) shows that the BBB dataset in MoleculeNet (Wu et al., 2018) has data curation errors, including duplicate structures with different labels. Therefore, AI agents will simulate humans by applying nuanced scientific intuition, weighing the strength of the evidence, and forming a reasoned judgment with limited data.

Hallucination and Reliability. AI models, particularly LLMs, often hallucinate plausible but false information (Li et al., 2024c; Sui et al., 2024), undermining their scientific credibility by fabricating or misrepresenting outcomes. For example, Sahoo et al. (2024) shows foundation models hallucination across modalities, including text, image, video, and audio. They also struggle to reason logically about complex biomedical and chemical tasks and are susceptible to simplistic biases or positional fallacies (DeLong et al., 2024; Joshi et al., 2024). Existing AI agents utilize external tools to address this problem. However, orchestrating tools reliably is non-trivial. For example, Yu et al. (2025) shows that tool augmentation does not always help chemistry questions. Therefore, we need to develop AI agents that can engage scientists in bidirectional, real-time collaboration, where dialogue, idea generation, and experimental design flow seamlessly between humans and machines.

Evaluation and Safety Evaluation is problematic, since their reasoning procedure and outputs might contain modalities other than text, making traditional text-based evaluation inadequate (Abramson et al., 2022). Furthermore, the vast, opaque training corpora of LLMs create a significant risk of data contamination (Magar and Schwartz, 2022). For example, Golchin and Surdeanu (2024) shows that GPT-4 exhibits significant data contamination across all evaluated datasets. Finally, AI safety concerns have intensified, highlighted by vulnerabilities like adversarial manipulation of medical language models and biosecurity risks posed by models generating harmful biological designs (Yang et al., 2024b; Brent and McKelvey Jr, 2025). Robust governance frameworks emphasizing truthfulness, resilience, fairness, robustness, and privacy are urgently required (Yang et al., 2024c).

Robotic Laboratory Automation An AI-generated experimental plan may be theoretically sound but practically infeasible due to physical factors. Physical lab setups often impose equipment limitations. Additionally, a more common obstacle is the software integration gap due to fragmented laboratory software that is difficult to scale or adapt (Scitara, 2023). As a result, current autonomous labs are limited to certain reactions, including solid-state synthesis(Szymanski et al., 2023), palladium-catalyzed cross-couplings (Boiko et al., 2023), and the synthesis of known small molecules (M. Bran et al., 2024), where hardware, software, and procedures have been tightly controlled and standardized. Future research can focus on creating agents that can generate robust, real-world experimental protocols that account for these practical constraints.

8 Limitation

In this survey, we present a comprehensive review of biomedical and chemical multimodal agents. However, due to the rapid evolution of this domain, a small portion of emergent methods and datasets may have been overlooked, especially those published close to our submission deadline. Additionally, due to space limitations, we acknowledge that there are areas closely related to our survey that have not been discussed in depth. For instance, diffusion-based models for drug discovery, such as ALIDIFF (Gu et al., 2024) and ABDPO (Zhou et al., 2024b), are not discussed in depth, though they represent a growing and important line of work. Furthermore, some AI Agents may belong to more than one category, given our classification criteria in the paper. For example, Agentic-Tx (Wang et al., 2025a) belongs to the ReAct framework (Yao et al., 2023) and instruction tuning.

Ethic Consideration

This study is a literature-based survey synthesizing publicly available information from previously published academic works. It does not involve human participants, animal subjects, newly collected datasets, or the use of private or identifiable data. All datasets and tools reviewed in this work were obtained from open-access sources. While we acknowledge that AI agents in biomedical and chemical domains may raise ethical concerns, this work itself does not pose direct ethical risks. Therefore, institutional ethical approval was not required.

Acknowledgment

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH) and by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the National Library of Medicine. The contributions of the NIH author(s) are considered Works of the United States Government. ORISE is managed by ORAU under DOE contract number DESC0014664. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of NIH, NLM, DOE, the U.S. Department of Health and Human Services, or ORAU/ORISE.

References

Naafey Aamer, Muhammad Nabeel Asim, Shan Munir, and Andreas Dengel. 2025. Automating ai discovery for biomedicine through knowledge graphs and llm agents. *bioRxiv*, pages 2025–05.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. 2024. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500.

Josh Abramson, Arun Ahuja, Federico Carnevale, Petko Georgiev, Alex Goldin, Alden Hung, Jessica Landon, Timothy Lillicrap, Alistair Muldal, Blake Richards, et al. 2022. Evaluating multimodal interactive agents. *Machine Learning Repository*, arXiv:2205.13274.

Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. 2022. Multimodal biomedical ai. *Nature medicine*, 28(9):1773–1784.

Abhinav Aggarwal. 2025. Rethinking llm benchmarks for 2025: Why agentic ai needs a new evaluation standard. *Fluid AI GPT - Enterprise GPT Solution*.

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Can knowledge graphs reduce hallucinations in LLMs? : A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3947–3960, Mexico City, Mexico. Association for Computational Linguistics.

Mariam ALMutairi, Lulwah AlKulaib, Melike Aktas, Sara Alsalamah, and Chang-Tien Lu. 2024. Synthetic Arabic medical dialogues using advanced multi-agent LLM techniques. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 11–26, Bangkok, Thailand. Association for Computational Linguistics.

Mehrad Ansari and Seyed Mohamad Moosavi. 2024. Agent-based learning of materials datasets from the scientific literature. *Digital Discovery*, 3(12):2607–2617.

Anthropic. 2024. Introducing the next generation of claude.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738, Albuquerque, New Mexico. Association for Computational Linguistics.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina

Salvatelli, Harshita Sharma, et al. 2024. Maira-2: Grounded radiology report generation. *Computation and Language Repository*, arXiv:2406.04449.

Hritik Bansal, Daniel Israel, Siyan Zhao, Shufan Li, Tung Nguyen, and Aditya Grover. 2025. Medmax: Mixed-modal instruction tuning for training biomedical assistants. *Computing Research Repository*, arXiv:2412.12661. Version 2.

Adrian Barnett and Zoe Doubleday. 2020. The growth of acronyms in the scientific literature. *elife*, 9:e60080.

Gonzalo Benegas, Chengzhong Ye, Carlos Albors, Jianan Canal Li, and Yun S Song. 2025. Genomic language models: opportunities and challenges. *Trends in Genetics*.

A Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J Bellis, Marleen De Veij, and Andrew R Leach. 2020. An open source chemical structure curation pipeline using rdkit. *Journal of Cheminformatics*, 12:1–16.

UK Biobank. 2014. About uk biobank.

Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O'Donovan, Isabelle Phan, et al. 2003. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31(1):365–370.

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

Albert Bou, Morgan Thomas, Sebastian Dittert, Carles Navarro, Maciej Majewski, Ye Wang, Shivam Patel, Gary Tresadern, Mazen Ahmad, Vincent Moens, et al. 2024. Acegen: Reinforcement learning of generative chemical agents for drug discovery. *Journal of Chemical Information and Modeling*, 64(15):5900–5911.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Roger Brent and T Greg McKelvey Jr. 2025. Contemporary ai foundation models increase biological weapons risk. *Computers and Society Repository*, arXiv:2506.13798.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, et al. 2025. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19552–19564.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025a. Unleashing the potential of prompt engineering for large language models. *Patterns*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, et al. 2025b. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159.

Xiaoyang Chen, Keyi Li, Xuejian Cui, Zian Wang, Qun Jiang, Jiacheng Lin, Zhen Li, Zijing Gao, and Rui Jiang. 2024. Epiagent: Foundation model for single-cell epigenomic data. *bioRxiv*, pages 2024–12.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR.

Bernardo P. de Almeida, Guillaume Richard, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Chandana Rajesh, Marie Lopez, Alexandre Laterre, Maren Lang, Uğur Şahin, Karim Beguir, and Thomas Pierrot. 2025. A multimodal conversational agent for dna, rna and protein tasks. *Nature Machine Intelligence*.

Nima Dehghani and Michael Levin. 2024. Bio-inspired ai: Integrating biological complexity into artificial intelligence. *Neurons and Cognition Repository*, arXiv:2411.15243.

Lauren Nicole DeLong, Yojana Gadiya, Paola Galdi, Jacques D Fleuriot, and Daniel Domingo-Fernández. 2024. Mars: A neurosymbolic approach for interpretable drug discovery. *Artificial Intelligence Repository*, arXiv:2410.05289.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. 2024. Agent ai: Surveying the horizons of multimodal interaction. *Artificial Intelligence Repository*, arXiv:2401.03568.

Kawin Ethayarajh and Dan Jurafsky. 2022. The authenticity gap in human evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6070, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. 2025. Medrax: Medical reasoning agent for chest x-ray. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Yongqi Fan, Kui Xue, Zelin Li, Xiaofan Zhang, and Tong Ruan. 2025a. An LLM-based framework for biomedical terminology normalization in social media via multi-agent collaboration. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10712–10726, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025b. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.

Bowen Gao, Yanwen Huang, Yiqiao Liu, Wenxuan Xie, Wei-Ying Ma, Ya-Qin Zhang, and Yanyan Lan. 2025. Pharmagents: Building a virtual pharma with large language model agents. *Biomolecules Repository*, arXiv:2503.22164.

Bowen Gao, Bo Qiang, Haichuan Tan, Yinjun Jia, Minsi Ren, Minsi Lu, Jingjing Liu, Wei-Ying Ma, and Yanyan Lan. 2023. Drugclip: Contrastive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems*, 36:44595–44614.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151.

Juan Jose Garau-Luis, Patrick Bordes, Liam Gonzalez, Maša Roller, Bernardo de Almeida, Christopher Blum, Lorenz Hexemer, Stefan Laurent, Maren Lang, Thomas Pierrot, and Guillaume Richard. 2024. Multi-modal transfer learning between biological foundation models. In *Advances in Neural Information Processing Systems*, volume 37, pages 78431–78450. Curran Associates, Inc.

Alireza Ghafarollahi and Markus J Buehler. 2024. Protagents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 3(7):1389–1409.

Alireza Ghafarollahi and Markus J. Buehler. 2025. Automating alloy design and discovery with physics-aware multimodal multiagent ai. *Proceedings of the National Academy of Sciences*, 122(4):e2414074122.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Jon M Laurent, Muhammed T Razzak, Andrew D White, Michaela M Hinks, and Samuel G Rodriques. 2025. Robin: A multi-agent system for automating scientific discovery. *Artificial Intelligence Repository*, arXiv:2505.13400.

GitHub. 2021. GitHub Copilot. `https://github.com/features/copilot`. Accessed: 2025-06-11.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, page 25–30. Association for Computing Machinery.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. 2025. Towards an ai co-scientist. *Artificial Intelligence Repository*, arXiv:2502.18864.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *Computation and Language Repository*, arXiv:2407.21783.

Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. 2023. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25.

Mourad Gridach, Jay Nanavati, Christina Mack, Khaldoun Zine El Abidine, and Lenon Mendes. 2025. Agentic ai for scientific discovery: A survey of progress, challenges, and future directions. In *Towards Agentic AI for Science: Hypothesis Generation, Comprehension, Quantification, and Validation*.

Siyi Gu, Minkai Xu, Alexander Powers, Weili Nie, Tomas Geffner, Karsten Kreis, Jure Leskovec, Arash Vahdat, and Stefano Ermon. 2024. Aligning target-aware molecule diffusion models with exact energy optimization. *Advances in Neural Information Processing Systems*, 37:44040–44063.

Valerio Guarrasi, Fatih Aksu, Camillo Maria Caruso, Francesco Di Feola, Aurora Rofena, Filippo Ruffini, and Paolo Soda. 2025. A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. *Image and Vision Computing*, 158:105509.

Kairui Guo, Mengjia Wu, Zelia Soo, Yue Yang, Yi Zhang, Qian Zhang, Hua Lin, Mark Grosser, Deon Venter, Guangquan Zhang, et al. 2023. Artificial intelligence-driven biomedical genomics. *Knowledge-Based Systems*, 279:110937.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Galib Muhammad Shahriar Himel, Md. Shourov Hasan, Umme Sadia Salsabil, and Md. Masudul Islam. 2024. Medlingua: A conceptual framework for a multilingual medical conversational agent. *MethodsX*, 12:102614.

Hongru Hu and Gerald Quon. 2024. scpair: Boosting single cell multimodal analysis by leveraging implicit feature selection and single cell atlases. *Nature Communications*, 15(1):9932.

Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. 2024. Crispr-gpt: An llm agent for automated design of gene-editing experiments. *Computing Research Repository*, arXiv:2404.18021.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. 2021. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. 2022. Artificial intelligence foundation for therapeutic science. *Nature chemical biology*, 18(10):1033–1036.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Junze Zhang, Yin Di, Shruti Marwaha, Jennefer Carter, Xin Zhou, Matthew T Wheeler, Jon Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael Snyder, Le Cong, Aviv Regev, and Jure Leskovec. 2025a. Biomni: A general-purpose biomedical ai agent. *bioRxiv*.

Qiuyuan Huang, Naoki Wake, Bidipta Sarkar, Zane Durante, Ran Gong, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Noboru Kuno, Ade Famoti, et al. 2025b. Position paper: Agent ai towards a holistic intelligence. *Artificial Intelligence Repository*, arXiv:2403.00833.

Hazel Inskip, Georgia Ntani, Leo Westbury, Chiara Di Gravio, Stefania D'Angelo, Camille Parsons, and Janis Baird. 2017. Getting started with tables. *Archives of Public Health*, 75:1–10.

Yan A. Ivanenkov, Daniil Polykovskiy, Dmitry Bezrukov, Bogdan Zagribelnyy, Vladimir Aladinskiy, Petrina Kamya, Alex Aliper, Feng Ren, and Alex Zhavoronkov. 2023. Chemistry42: An ai-driven platform for molecular design and optimization. *Journal of Chemical Information and Modeling*, 63(3):695–701. PMID: 36728505.

Jiyue Jiang, Zikang Wang, Yuheng Shan, Heyan Chai, Jiayi Li, Zixian Ma, Xinrui Zhang, and Yu Li. 2025. Biological sequence with language model prompting: A survey. *Computation and Language Repository*, arXiv:2503.04135.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, W John Wilbur, Zhe He, Andrew Taylor, Qingyu Chen, et al. 2024a. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *Computation and Language Repository*, arXiv:2402.13225.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2024b. Genegpt: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. 2024. LLMs are prone to fallacies in causal inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10553–10569, Miami, Florida, USA. Association for Computational Linguistics.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Yeonghun Kang and Jihan Kim. 2024. Chatmof: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nature communications*, 15(1):4705.

Hyomin Kim, Yunhui Jang, and Sungsoo Ahn. 2025. Mt-mol: Multi agent system with tool-based reasoning for molecular optimization. *Artificial Intelligence Repository*, arXiv:2505.20820.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452. Curran Associates, Inc.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

Esther Landhuis. 2016. Scientific literature: Information overload. *Nature*, 535(7612):457–458.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. 2024. Lab-bench: Measuring capabilities of language models for biology research. *Artificial Intelligence Repository*, arXiv:2407.10362.

Wanlu Lei, Caterina Fuster-Barceló, Gabriel Reder, Arrate Muñoz-Barrutia, and Wei Ouyang. 2024. Bioimage. io chatbot: a community-driven ai assistant for integrative computational bioimaging. *nature methods*, 21(8):1368–1370.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. 2024a. MMedAgent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760, Miami, Florida, USA. Association for Computational Linguistics.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564. Curran Associates, Inc.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. 2024b. Agent hospital: A simulacrum of hospital with evolvable medical agents. *Computing Research Repository*, arXiv:2405.02957.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024c. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

Zhuoqun Li, Hongyu Lin, Yaojie Lu, Hao Xiang, Xianpei Han, and Le Sun. 2024d. Meta-cognitive analysis: Evaluating declarative and procedural knowledge in datasets and large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11222–11228, Torino, Italia. ELRA and ICCL.

David J Lipman and William R Pearson. 1985. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441.

Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Yingzhou Lu, and Yue Zhao. 2024. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *Machine Learning Repository*, arXiv:2411.15692.

Yungeng Liu, Zan Chen, Yuguang Wang, and Yiqing Shen. 2025. AutoProteinEngine: A large language model driven agent framework for multimodal AutoML in protein engineering. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 422–430, Abu Dhabi, UAE. Association for Computational Linguistics.

Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, et al. 2025. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19724–19735.

Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024a. TriageAgent: Towards better multi-agents collaborations for large language model-based clinical triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5747–

5764, Miami, Florida, USA. Association for Computational Linguistics.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. 2024b. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473.

Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. 2024c. ClinicalRAG: Enhancing clinical decision support through heterogeneous knowledge retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 64–68, Bangkok, Thailand. Association for Computational Linguistics.

Li Lucy, Jesse Dodge, David Bamman, and Katherine Keith. 2023. Words as gatekeepers: Measuring discipline-specific terms and meanings in scholarly publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6929–6947, Toronto, Canada. Association for Computational Linguistics.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. Llm4sr: A survey on large language models for scientific research. *Artificial Intelligence Repository*, arXiv:2501.04306.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11.

Kangyong Ma. 2025. Ai agents in chemical research: Gvim–an intelligent research assistant system. *Digital Discovery*.

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. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

M. Maruf, Arka Daw, Kazi Sajeed Mehrab, Harish Babu Manogaran, Abhilash Neog, Medha Sawhney, Mridul Khurana, James Balhoff, Yasin Bakis, Bahadir Altintas, Matthew Thompson, Elizabeth Campolongo, Josef Uyeda, Hilmar Lapp, Henry Bart, Paula Mabee, Yu Su, Wei-Lun (Harry) Chao, Charles Stewart, Tanya Berger-Wolf, Wasila Dahdul, and Anuj Karpatne. 2024. Vlm4bio: A benchmark dataset to evaluate pretrained vision-language models for trait discovery from biological images. In *Advances in Neural Information Processing Systems*, volume 37, pages 131035–131071. Curran Associates, Inc.

Nicholas Matsumoto, Hyunjun Choi, Jay Moran, Miguel E Hernandez, Mythreye Venkatesan, Xi Li, Jui-Hsuan Chang, Paul Wang, and Jason H Moore. 2025. Escargot: an ai agent leveraging large language models, dynamic graph of thoughts, and biomedical knowledge graphs for enhanced reasoning. *Bioinformatics*, 41(2):btaf031.

Andrew D. McNaughton, Gautham Krishna Sankar Ramalaxmi, Agustin Kruel, Carter R. Knutson, Rohith A. Varikoti, and Neeraj Kumar. 2024. Cactus: Chemistry agent connecting tool usage to science. *ACS Omega*, 9(46):46563–46573.

Naomi Miller, Eve-Marie Lacroix, and Joyce EB Backus. 2000. Medlineplus: building and maintaining the national library of medicine's consumer health web service. *Bulletin of the Medical Library Association*, 88(1):11.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, et al. 2025. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, pages 1–8.

Ludovico Mitchener, Jon M Laurent, Benjamin Tenmann, Siddharth Narayanan, Geemi P Wellawatte, Andrew White, Lorenzo Sani, and Samuel G Rodriques. 2025. Bixbench: a comprehensive benchmark for llm-based agents in computational biology. *Quantitative Methods Repository*, arXiv:2503.00096.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Damien Olivier-Jimenez, Rico J. E. Derks, Oscar Harari, Carlos Cruchaga, Muhammad Ali, Alessandro Ori, Domenico Di Fraia, Birol Cabukusta, Andy Henrie, Martin Giera, and Yassene Mohammed. 2025. isoda: A comprehensive tool for integrative omics data analysis in single- and multi-omics experiments. *Analytical Chemistry*, 97(5):2689–2697. PMID: 39886798.

OpenAI. 2023. Gpt-4 technical report. *Computation and Language Repository*, arXiv:2303.08774.

Zetian Ouyang, Yishuai Qiu, Linlin Wang, Gerard De Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. CliMedBench: A large-scale Chinese benchmark for evaluating medical large language models in clinical scenarios. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8428–8438, Miami, Florida, USA. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Himanshu Gautam Pandey, Akhil Amod, and Shivang Kumar. 2024. Advancing healthcare automation: Multi-agent system for medical necessity justification. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 39–49, Bangkok, Thailand. Association for Computational Linguistics.

Hyunsoo Park, Sauradeep Majumdar, Xiaoqi Zhang, Jihan Kim, and Berend Smit. 2024. Inverse design of metal–organic frameworks for direct air capture of co 2 via deep reinforcement learning. *Digital Discovery*, 3(4):728–741.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. 2024. BioT5+: Towards generalized biological understanding with IUPAC integration and multi-task tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1216–1240, Bangkok, Thailand. Association for Computational Linguistics.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. BioT5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1123, Singapore. Association for Computational Linguistics.

Edward O Pyzer-Knapp, Jed W Pitera, Peter WJ Staar, Seiji Takeda, Teodoro Laino, Daniel P Sanders, James Sexton, John R Smith, and Alessandro Curioni. 2022. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials*, 8(1):84.

Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. 2024. Large language models as biomedical hypothesis generators: A comprehensive evaluation. In *First Conference on Language Modeling*.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2025. Tool learning with large language models: A survey. *Frontiers of Computer Science*, 19(8):198343.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *OpenAI*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Mayk Caldas Ramos, Christopher J Collison, and Andrew D White. 2025. A review of large language models and autonomous agents in chemistry. *Chemical Science*.

Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. 2025. Towards scientific intelligence: A survey of llm-based scientific agents. *Artificial Intelligence Repository*, arXiv:2503.24047.

Joseph D Romano, Van Truong, Rachit Kumar, Mythreye Venkatesan, Britney E Graham, Yun Hao, Nick Matsumoto, Xi Li, Zhiping Wang, Marylyn D Ritchie, et al. 2024. The alzheimer's knowledge base: A knowledge graph for alzheimer disease research. *Journal of Medical Internet Research*, 26:e46777.

Yusuf H Roohani, Jian Vora, Qian Huang, Percy Liang, and Jure Leskovec. 2024. Biodiscoveryagent: An AI agent for designing genetic perturbation experiments. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Loïc A Royer. 2024. Omega—harnessing the power of large language models for bioimage analysis. *Nature Methods*, pages 1–3.

Yixiang Ruan, Chenyin Lu, Ning Xu, Yuchen He, Yixin Chen, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, et al. 2024. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nature communications*, 15(1):10160.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

Daan Schouten, Giulia Nicoletti, Bas Dille, Catherine Chia, Pierpaolo Vendittelli, Megan Schuurmans, Geert Litjens, and Nadieh Khalili. 2025. Navigating the landscape of multimodal ai in medicine: A scoping review on technical challenges and clinical applications. *Medical Image Analysis*, page 103621.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Computation and Language Repository*, arXiv:1707.06347.

Scitara. 2023. Streamlining laboratory operations: Overcoming the challenges of point-to-point integrations.

Douglas W Selinger, Timothy R Wall, Eleni Stylianou, Ehab M Khalil, Jedidiah Gaetz, and Oren Levy. 2024. A framework for autonomous ai-driven drug discovery. *bioRxiv*, pages 2024–12.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C. Ho, Carl Yang, and May Dongmei Wang. 2024. EHRAgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22315–22339, Miami, Florida, USA. Association for Computational Linguistics.

Khachik Smbatyan, Tsolak Ghukasyan, Tigran Aghajanyan, Hovhannes Dabaghyan, Sergey Adamyan, Aram Bughdaryan, Vahagn Altunyan, Gagik Navasardyan, Aram Davtyan, Anush Hakobyan, et al. 2025. Can ai agents design and implement drug discovery pipelines? *Artificial Intelligence Repository*, arXiv:2504.19912.

Vivek Sriram, Ashley Mae Conard, Ilyana Rosenberg, Dokyoon Kim, T Scott Saponas, and Amanda K Hall. 2025. Addressing biomedical data challenges and opportunities to inform a large-scale data lifecycle for enhanced data sharing, interoperability, analysis, and collaboration across stakeholders. *Scientific Reports*, 15(1):6291.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. 2025. Paperbench: Evaluating ai's ability to replicate ai research. *Artificial Intelligence Repository*, arXiv:2504.01848.

Isabella Stewart and Markus J Buehler. 2025. Molecular analysis and design using generative artificial intelligence via multi-agent modeling. *Molecular Systems Design & Engineering*, 10(4):314–337.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard So. 2024. Confabulation: The surprising value of large language model hallucinations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14274–14284, Bangkok, Thailand. Association for Computational Linguistics.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Yu Rong, Wenbing Huang, Qifeng Bai, and Tingyang Xu. 2025. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. volume arXiv:2506.09513.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2024. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, pages 2024–11.

Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. 2023. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *Computation and Language Repository*, arXiv:2408.00118.

Brian H Toby and Robert B Von Dreele. 2013. Gsas-ii: the genesis of a modern open-source all purpose crystallography software package. *Applied Crystallography*, 46(2):544–549.

Gary Tom, Stefan P Schmid, Sterling G Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M Rajaonson, Marta Skreta, et al. 2024. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732.

Susana M Vieira, Uzay Kaymak, and João MC Sousa. 2010. Cohen's kappa coefficient as a performance measure for feature selection. In *International conference on fuzzy systems*, pages 1–8. IEEE.

Anita Ioana Visan and Irina Negut. 2024. Integrating artificial intelligence for drug discovery in the context of revolutionizing drug delivery. *Life*, 14(2):233.

Pat Walters. 2023. We need better benchmarks for machine learning in drug discovery.

Eric Wang, Samuel Schmidgall, Paul F Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. 2025a. Txgemma: Efficient and agentic llms for therapeutics. *Artificial Intelligence Repository*, arXiv:2504.06196.

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

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024b. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299, Bangkok, Thailand. Association for Computational Linguistics.

Qingyun Wang, Zixuan Zhang, Hongxiang Li, Xuan Liu, Jiawei Han, Huimin Zhao, and Heng Ji. 2024c. Chem-FINESE: Validating fine-grained few-shot entity extraction through text reconstruction. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1–16, St. Julian's, Malta. Association for Computational Linguistics.

Ruheng Wang, Yi Jiang, Junru Jin, Chenglin Yin, Haoqing Yu, Fengsheng Wang, Jiuxin Feng, Ran Su, Kenta Nakai, Quan Zou, et al. 2023b. Deepbio: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis. *Nucleic acids research*, 51(7):3017–3029.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024d. CMB: A comprehensive medical benchmark in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205, Mexico City, Mexico. Association for Computational Linguistics.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024e. Executable code actions elicit better llm agents. In *Forty-first International Conference on Machine Learning*.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024f. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279, Mexico City, Mexico. Association for Computational Linguistics.

Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, and Zhiyong Lu. 2024g. Geneagent: Self-verification language agent for gene set knowledge discovery using domain databases. *Artificial Intelligence Repository*, arXiv:2405.16205.

Zifeng Wang, Benjamin Danek, and Jimeng Sun. 2025b. Biodsa-1k: Benchmarking data science agents for biomedical research. *Artificial Intelligence Repository*, arXiv:2505.16100.

Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen Harrison, Chengwei Pan, Junyi Gao, and Liantao Ma. 2025c. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2250–2261, New York, NY, USA. Association for Computing Machinery.

Ziyue Wang, Junde Wu, Chang Han Low, and Yueming Jin. 2025d. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *Artificial Intelligence Repository*, arXiv:2503.18968.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.

Eric Wu, Kevin Wu, and James Zou. 2025a. Medarena: Comparing llms for medicine in the wild. *Stanford HAI*.

Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. 2025b. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102:103547.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

Qi Xin, Quyu Kong, Hongyi Ji, Yue Shen, Yuqi Liu, Yan Sun, Zhilin Zhang, Zhaorong Li, Xunlong Xia, Bing Deng, et al. 2024. Bioinformatics agent (bia): Unleashing the power of large language models to reshape bioinformatics workflow. *bioRxiv*, pages 2024–05.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025:*

Proceedings of the Pacific Symposium, pages 199–214. World Scientific.

Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Xiangru Tang, Hang Wu, May D Wang, Peifeng Ruan, Donghan Yang, Tao Wang, et al. 2025. Medagentgym: Training llm agents for code-based medical reasoning at scale. *Computation and Language Repository*, arXiv:2506.04405.

Chih-Hsuan Yang, Benjamin Feuer, Talukder "Zaki" Jubery, Zi Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh Singh, ARTI SINGH, Soumik Sarkar, Nirav Merchant, Chinmay Hegde, and Baskar Ganapathysubramanian. 2024a. Biotrove: A large curated image dataset enabling ai for biodiversity. In *Advances in Neural Information Processing Systems*, volume 37, pages 102101–102120. Curran Associates, Inc.

Yifan Yang, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024b. Adversarial attacks on large language models in medicine. *Artificial Intelligence Repository*, arXiv:2406.12259.

Yifan Yang, Qiao Jin, Robert Leaman, Xiaoyu Liu, Guangzhi Xiong, Maame Sarfo-Gyamfi, Changlin Gong, Santiago Ferrière-Steinert, W John Wilbur, Xiaojun Li, et al. 2024c. Ensuring safety and trust: Analyzing the risks of large language models in medicine. *Computation and Language Repository*, arXiv:2411.14487.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Yanbin Yin, Zhen Wang, Kun Zhou, Xiangdong Zhang, Shibo Hao, Yi Gu, Jieyuan Liu, Somanshu Singla, Tianyang Liu, Eric P. Xing, Zhengzhong Liu, Haojian Jin, and Zhiting Hu. 2024. Decentralized arena via collective llm intelligence: Building automated, robust, and transparent llm evaluation for numerous dimensions.

Nicholas D Youngblut, Christopher Carpenter, Jaanak Prashar, Chiara Ricci-Tam, Rajesh Ilango, Noam Teyssier, Silvana Konermann, Patrick D Hsu, Alexander Dobin, David P Burke, et al. 2025. scbasecount: an ai agent-curated, uniformly processed, and continually expanding single cell data repository. *bioRxiv*, pages 2025–02.

Botao Yu, Frazier N. Baker, Ziru Chen, Garrett Herb, Boyu Gou, Daniel Adu-Ampratwum, Xia Ning, and Huan Sun. 2025. Tooling or not tooling? the impact of tools on language agents for chemistry problem solving. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7620–7640, Albuquerque, New Mexico. Association for Computational Linguistics.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. 2024. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Hu Jinfang, Zhiyuan Liu, and Bowen Zhou. 2024a. Ultramedical: Building specialized generalists in biomedicine. In *Advances in Neural Information Processing Systems*, volume 37, pages 26045–26081. Curran Associates, Inc.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. 2025. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024b. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8783–8817, Miami, Florida, USA. Association for Computational Linguistics.

Zuobai Zhang, Minghao Xu, Arian Rokkum Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. 2023. Protein representation learning by geometric structure pretraining. In *The Eleventh International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. *Computation and Language Repository*, arXiv:2505.13259.

Juexiao Zhou, Bin Zhang, Guowei Li, Xiuying Chen, Haoyang Li, Xiaopeng Xu, Siyuan Chen, Wenjia He, Chencheng Xu, Liwei Liu, et al. 2024a. An ai agent

for fully automated multi-omic analyses. *Advanced Science*, 11(44):2407094.

Xiangxin Zhou, Dongyu Xue, Ruizhe Chen, Zaixiang Zheng, Liang Wang, and Quanquan Gu. 2024b. Antigen-specific antibody design via direct energy-based preference optimization. *Advances in Neural Information Processing Systems*, 37:120861–120891.

Yakun Zhu, Shaohang Wei, Xu Wang, Kui Xue, Shaoting Zhang, and Xiaofan Zhang. 2025. MeNTi: Bridging medical calculator and LLM agent with nested tool calling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5097–5116, Albuquerque, New Mexico. Association for Computational Linguistics.

A Related Work

Scientific discovery is an inherently multimodal process, involving data formats beyond text, such as molecular structures, clinical images, protein interaction networks, and biomedical sensor signals. Foundational paper (Durante et al., 2024) introduces multimodal agentic AI as systems that perceive diverse environmental modalities and execute embodied actions. While their work focuses primarily on classical modalities such as language, vision, and action, our study broadens the scope to include domain-specific modalities central to biomedicine and chemistry. In parallel, Wang et al. (2024a) provides a general overview of LLM-based agent frameworks and their applications, but remains domain-agnostic. Recently, researchers are increasingly interested in AI agents designed for scientific discovery. Luo et al. (2025) and (Zheng et al., 2025) offer a comprehensive examination of how LLMs are being integrated across the entire scientific research pipeline in general science domains, from hypothesis discovery and experiment planning through writing and peer review. Compared to it, Ren et al. (2025); Gridach et al. (2025) survey architectural taxonomies and system-level insights into scientific AI agents. Following Durante et al. (2024), (Ren et al., 2025) divides the agent architecture into three core modules: a Planner for task decomposition, a Memory system for context retention and learning, and a Tool Set for executing actions. Compared to us, it only covers some papers about AI agents in biomedical and chemical domains, and most of its papers talk about general scientific research AI agents. Gridach et al. (2025) discusses existing Agentic AI systems, frameworks, and methods in chemistry, biology, and materials science. However, it fails to provide detailed guidance on how researchers can practically integrate Agentic AI into their workflows. On the position side, Gao et al. (2024) presents a vision of "AI scientists" as collaborative partners that combine human creativity with AI's ability to analyze large datasets and navigate vast hypothesis spaces in the biomedical domain by integrating with experimental platforms, planning, and evaluating experiments, and leveraging human-AI collaboration. However, their work emphasizes vision over concrete implementation or domain applications. Ramos et al. (2025) covers LLM-driven agents applied to chemistry, highlighting document processing, synthesis planning, and lab tool integration. While it reflects on challenges such as data quality, interpretability, and benchmarking, its scope remains focused on chemistry and materials science. Building on this prior work, our paper brings together developments in AI agents across both biomedical and chemical domains, two areas that have rarely been studied in combination. We take a deep dive into the unique data types (modalities), learning methods, reasoning abilities, and real-world applications of multimodal agentic AI in these fields. To support this, we curated and analyzed over 80 relevant papers, providing a comprehensive overview of how these agents are being developed, utilized, and evaluated in practice. We summarize the aims and domains of these and other recent survey papers in Table E.12.

B Multimedia Modalities Covered by Biomedical and Chemical Research

Sequential Data + Text. Sequential encoding is essential in the development of biomedical and chemical AI agents, since chemical sequences (i.e., DNA, RNA, and proteins) provide complementary chemical properties that are difficult to capture in descriptive text. Recent research is increasingly emphasizing omics sequence data over biomedical text (Acosta et al., 2022; Gao et al., 2024; Benegas et al., 2025), driven by the growing availability of large-scale omics datasets (Boeckmann et al., 2003; Biobank, 2014; Grešová et al., 2023) and the structural properties of biological sequences (Wang et al., 2023b; Guo et al., 2023; Abramson et al., 2024). A common strategy is to steer LLMs by natural language in analyzing biological sequences (Jiang et al., 2025). For example, ChatNT (de Almeida et al., 2025) uses En-

glish prompts to perform 27 genomics-related tasks, while AutoProteinEngine (Liu et al., 2025) applies LLMs to automate protein engineering. Compared to previous multimodal foundation models (Pei et al., 2023; Zhang et al., 2023; Garau-Luis et al., 2024; Pei et al., 2024; Chen et al., 2024), both agents offer a conversational interface that is particularly useful for users with no coding capabilities.

Visual Data + Text. Images provide visual representations of biological structures and processes. Existing papers primarily focus on medical reports and figure-caption pairs extracted from papers (Zhang et al., 2024b). Most biomedical vision-language agents leverage GPT-4V (OpenAI, 2023), retrieval-augmented generation (RAG) (Lewis et al., 2020), and prompt engineering (Chen et al., 2025a) for agent-driven conversational bioimage analysis (Royer, 2024; Lei et al., 2024). In addition to GPT-4V, some papers enhance domain-specific performance by instruction-tuning open-source vision-language models (Lu et al., 2024b; Li et al., 2023) for human pathology analysis, which aligns the models more closely with biomedical tasks by using curated prompts and task-specific training objectives. For example, MMedAgent (Li et al., 2024a), PathChat (Lu et al., 2024b), and MedMax (Bansal et al., 2025), have demonstrated superior performance compared to GPT-4 on multiple medical tasks by combining specialized medical vision encoders with large language models.

Structured Data + Text. Structured data transforms complex datasets into organized, accessible formats. In the biomedical and chemical domains, structured data can be divided into several types, including tabular data (Inskip et al., 2017; Sriram et al., 2025), knowledge bases (Chandak et al., 2023), and Omics Data (Olivier-Jimenez et al., 2025). Tabular data are organized in rows and columns, which are used for storing clinical measurements, electronic health records (Wang et al., 2025c; Shi et al., 2024; Zhu et al., 2025), laboratory test results (Boiko et al., 2023), and chemical properties (M. Bran et al., 2024). AI agents for other applications also benefit from external knowledge bases (Ansari and Moosavi, 2024; Wang et al., 2024b; Selinger et al., 2024; Wang et al., 2024g; Aamer et al., 2025; Lu et al., 2024c; Matsumoto et al., 2025), by strengthening the reasoning process with domain knowledge, which provides more accurate, explainable, and context-aware responses. For example, Intelliscope (Aamer et al., 2025) and GeneAgent (Wang et al., 2024g) combine LLMs with structured biological databases to improve scientific hypothesis generation and gene set analysis. Omics datasets are large-scale, comprehensive datasets that capture various layers of biological information in genes, proteins, transcripts, etc. Previous analysis of different omics datasets requires professional skills and domain knowledge. To solve this problem, Automated Bioinformatics Analysis (Zhou et al., 2024a) fully automates multi-omic analyses based on LLMs through code generation. Additionally, Youngblut et al. (2025) proposes a hierarchical multi-agent collaborative pipeline to discover datasets and curate metadata from the Sequence Read Archive automatically.

C Additional Strategies for Multimodal Agentic AI Learning

C.1 Domain-specific Tuning (Table E.4)

To adapt models to a particular field, researchers usually perform domain-specific tuning. Gururangan et al. (2020) shows that task-adaptive pretraining on a smaller but task-relevant corpus can boost performance.

Instruction Tuning. Instruction tuning, also known as supervised finetuning (SFT), is a prominent sub-category where models are finetuned on datasets of instruction–output pairs. The goal is to infuse models with domain-specific knowledge and enhance their ability to follow instructions effectively. MMedAgent (Li et al., 2024a), for instance, is trained on a curated instruction-tuning dataset that includes six medical tools designed to solve seven tasks across five different modalities, which explicitly teaches the agent to choose the most suitable tools for a given medical task. A more extensive example is TxGemma (Wang et al., 2025a), which finetunes the Gemma-2 base models (Team et al., 2024) on a comprehensive dataset from the Therapeutic Data Commons (TDC) (Huang et al., 2021, 2022), covering 66 AI-ready drug discovery datasets formatted as prompts with an instruction, context, a question on therapeutic properties, and an answer. This large-scale instruction tuning, with approximately 7 million training examples, aims to create specialized therapeutic LLMs. Similarly, in the chemical domain, GVIM (Ma, 2025) equips the model with specialized chemical knowledge and reasoning by finetuning open-source LLMs, such as LLaMA-3 (Grattafiori et al., 2024), on instructional data collected from the field of chemical science.

Reinforcement Learning Instruction tuning enables AI agents to follow human directives, but it falls short in enhancing their ability to interact with external environments or autonomously explore and learn new tasks. In biomedical and chemical domains, task evaluation typically involves multiple abstract criteria (e.g., balancing immediate clinical benefits with long-term patient outcomes). To address this complexity, reinforcement learning with well-crafted reward functions empowers AI agents to iteratively interact with an environment and learn optimal policies. These reward functions guide agents toward desired behaviors by providing feedback over time, shaping their learning process to handle the nuanced, multi-objective nature of these tasks. For example, ACE-GEN (Bou et al., 2024) is a reinforcement learning toolkit for generative chemical agents in drug discovery based on drug design relevant scoring functions. Chemistry42 (Ivanenkov et al., 2023) uses multiple sets of reward modules to dynamically evaluate the properties of generated structures in 2D and 3D against user-defined criteria. Szymanski et al. (2023) trains agent with proximal policy optimization (Schulman et al., 2017), which can interact with GSAS-II software package (Toby and Von Dreele, 2013). ChatMOF (Kang and Kim, 2024) leverages reinforcement learning (Park et al., 2024) to select building blocks for metal-organic frameworks. Direct Preference Optimization (DPO) (Rafailov et al., 2023) is a reward-free finetuning strategy previously used in natural language processing.

D Additional Evaluation and Benchmarks

We summarize the recent benchmark papers in Table E.13.

D.1 Task-Oriented Evaluation

Task-specific Benchmarks. Task-specific benchmarks aim to assess an agent's ability to perform a complete, end-to-end task. For example, BioDiscoveryAgent (Roohani et al., 2024) defines "hits" as genes whose perturbation leads to a desired phenotype. Its performance is measured via Hit Rate (Recall), reflecting the proportion of true positive hits identified within a dataset. Similarly, Biomni (Huang et al., 2025a) is evaluated on a suite of eight newly curated, realistic biomedical tasks, including variant prioritization, GWAS causal gene detection, CRISPR perturbation screen design, rare disease diagnosis, patient gene prioritization, drug repurposing, microbiome disease-taxa analysis, and single-cell RNA-seq annotation. Evaluation metrics are tailored to each task, ranging from classification accuracy and average post-perturbation effect to semantic matching accuracy verified by human experts. In bioinformatics, BioDSA-1K (Wang et al., 2025b) provides 1,029 hypothesis-centric tasks paired with 1,177 analysis plans, curated from over 300 published biomedical studies to reflect the structure and reasoning found in authentic research workflows. It evaluates AI agents for the full research workflow, including hypothesis formulation, data analysis, evidence alignment, and code execution. Notably, it also includes non-verifiable cases, requiring agents to recognize when available data are insufficient to support or refute a claim. In the medical domain, AgentMD (Jin et al., 2024a) introduces a benchmark, RiskQA, specifically for evaluating an agent's ability to perform clinical risk prediction that assesses both tool creation (via quality-check accuracy and unit-test pass rate) and tool usage (via RiskQA task accuracy). Finally, TriageAgent (Lu et al., 2024a) provides the first publicly available benchmark dataset of notes annotated with Emergency Severity Index (ESI) levels and human-expert performance. The study evaluates model safety and accuracy using Discordance Rate as the primary metric, complemented by under-triage and over-triage rates.

Interactive Benchmarks. Recent evaluation methods focus on interactive benchmarks that require agents to operate over time. Those benchmarks assess not just the final outcome but the entire process of interaction, decision-making, and adaptation for AI agents. For example, Agent Hospital (Li et al., 2024b) simulates a fully functioning hospital, where LLM-powered agents play the roles of patients, nurses, and doctors. Instead of a static score, the evaluation focuses on how the doctor agent improves over time, regarding diagnostic accuracy, the appropriateness of selected medical tests, and the quality of treatment recommendations. Similarly, AI Hospital (Fan et al., 2025b) introduces the Multi-View Medical Evaluation (MVME) benchmark to assess LLM-driven "Doctor" agents in a simulated clinical environment. It computes entity overlap-based automated metrics for the Diagnostic Results section of medical records.

Real-world Evaluation. Real-world outcomes,

such as wet-lab experiments, represent the ultimate benchmark for testing AI agents. For example, in Biomni (Huang et al., 2025a), a scientist assigned Biomni an RNA cloning task and followed its protocol exactly to perform the wet-lab experiment. The results show how scientists can rely on Biomni to autonomously design complex molecular biology experiments with accuracy comparable to human experts. A more ambitious goal is to evaluate agents that interact with physical laboratory hardware, grounding evaluation in real-world outcomes by designing a Self-Driving Laboratory (SDL). In this context, key evaluation metrics are tightly coupled to scientific objectives, such as yield, purity, reaction mass efficiency, space-time yield, and E-factor (Tom et al., 2024). For example, Coscientist (Boiko et al., 2023) frames reaction optimization as a game aimed at maximizing reaction yield, measured through normalized advantage over iterative loops. Similarly, LLM-RDF (Ruan et al., 2024) tracks both yield and probability of improvement (PI) during closed-loop optimization to decide when experiments should stop. Meanwhile, ChemCrow (M. Bran et al., 2024) compares automated protocol performance against expert benchmarks in both task success rates and qualitative expert assessments.

D.2 Subjective Evaluation

Human/LLMs Ranking Existing benchmarks focus on scientific knowledge, which leads to a "benchmark mirage" (Aggarwal, 2025): the illusion that high performance on benchmarks translates directly to real-world effectiveness of AI agents. In reality, these benchmarks fall short of evaluating what truly defines an autonomous agent: the ability to operate over time, manage memory and internal state, interact with tools and APIs, and adapt continuously to changing environments. To solve this, researchers propose interactive arenas to test and compare top-performing AI agents. For example, MedArena (Wu et al., 2025a) is a free, interactive arena for clinicians to test and compare top-performing LLMs on their medical queries. Each annotator in the arena is required to provide a National Provider Identifier (NPI) or a Doximity account. The ranking algorithm is based on the Bradley–Terry (BT) model (Bradley and Terry, 1952). As of March 2025, MedArena had collected over 1,200 preferences from more than 300 clinicians across 11 top-performing LLMs. Due to the limited scalability of human annotations, researchers leverage the collective intelligence of LLMs to evaluate and compare themselves. For instance, Decentralized Arena (Yin et al., 2024) automates and scales Chatbot Arena (Chiang et al., 2024) for LLM evaluation across various fine-grained dimensions (e.g., biology, chemistry, ...). The proposed arena achieves 95% correlation to Chatbot Arena, which relies on extensive human judges.

Human Qualitative Evaluation For many complex agentic tasks, purely quantitative metrics are insufficient or non-existent. In these cases, evaluation relies on direct, structured judgment by human experts (Ethayarajh and Jurafsky, 2022; Starace et al., 2025). This human-in-the-loop approach provides the gold standard for assessing qualities such as scientific soundness, practical utility, and interpretability. For example, in Biomni (Huang et al., 2025a), blinded expert reviewers evaluate protocols generated by a LLM (Claude (Anthropic, 2024)), Biomni, a human trainee, and a senior human expert. This evaluation procedure enables an unbiased, expert-driven comparison of the agent's capabilities. GeneAgent (Wang et al., 2024g) uses manual review to confirm the effectiveness of the self-verification module. Similarly, in ChemCrow (M. Bran et al., 2024), four human evaluators directly compare its outputs to GPT-4's and indicate their preference.

LLM-as-Judge Relying only on human domain experts for evaluation is often expensive, time-consuming, and difficult to scale. At the same time, simple automated metrics are too crude to capture the details required for these assessments. To bridge this gap, researchers rely on LLM-as-a-Judge to assess the quality of AI-generated responses (Zheng et al., 2023). The Judge LLMs are usually guided by a carefully engineered framework of prompts, detailed rubrics, and specific evaluation criteria. For example, AI Hospital (Fan et al., 2025b) employs GPT-4 (OpenAI, 2023) as an evaluator to score diagnostic reports generated by the agent based on a predefined set of options on a discrete 1-4 scale. A parallel human evaluation shows less than a 4% difference from the GPT-4 scores. ResearchAgent (Baek et al., 2025) prompts GPT-4 with human-annotated examples to induce a detailed, 5-point Likert scale rubric for a variety of metrics, including Clarity, Relevance, Originality, Feasibility, and Significance. Similarly, LLM4BioHypoGen (Qi et al., 2024) uses GPT-4 to evaluate the quality of generated

hypotheses regarding novelty, relevance, significance, and verifiability. Both papers observe high human-model agreement based on Spearman's correlation coefficient (Zar, 2005) and Cohen's kappa coefficient (Vieira et al., 2010). In bioinformatics, Bixbench (Mitchener et al., 2025) offers 53 real-world analytical scenarios and 296 open-ended questions to evaluate the performance of AI agents in biological data analysis. The generation results for open-ended questions are then evaluated by Claude 3.5 Sonnet (Anthropic, 2024).

D.3 Evaluation Metrics

Traditional task-specific benchmarks typically rely on well-defined, quantitative metrics, including accuracy, precision, recall, F1 Score, and ROC-AUC. Since AI agents usually have planning ability, researchers are increasingly shifting their evaluation focus from just final outcomes to examining the process and behavior of the agent (Durante et al., 2024). The emergence of agents that use code as their primary mode of action has also introduced a powerful new dimension for evaluation: the correctness and efficiency of the code itself. Therefore, researchers can design evaluation metrics to separate an agent's procedural competence from its declarative knowledge (Li et al., 2024d). Procedural competence refers to the agent's ability to formulate valid plans or generate executable code, while declarative knowledge concerns the factual accuracy of the final outputs. Metrics such as code executability (Shi et al., 2024; Huang et al., 2025a), goal completion rate (Boiko et al., 2023; M. Bran et al., 2024), and resource efficiency (Kim et al., 2024) offer direct ways to evaluate procedural competence, independent of whether the end result is factually correct. Conversely, metrics like scientific validity (Wang et al., 2024g; Huang et al., 2025a), novelty (Wang et al., 2024b; Baek et al., 2025), and verifiability (ALMutairi et al., 2024) assess the quality and impact of the agent's final conclusions, reflecting its declarative knowledge. Additionally, behavioral metrics such as adaptability (Shi et al., 2024) and interpretability (Roohani et al., 2024) further contextualize an agent's procedural abilities, highlighting how it handles errors or communicates its reasoning. Together, these metrics provide a more nuanced and comprehensive framework for evaluating AI agents beyond traditional outcome-based benchmarks.

E Tables

Modality	Strengths	Limitations	Use Cases
Textual Descriptions	- Universally interpretable by LLMs - Directly compatible with natural language prompting - Supports integration with other modalities	- Ambiguity and incompleteness in clinical narratives affect reasoning accuracy - Requires external validation for reliability	Medical report summarization, question answering, guideline retrieval
Tabular Data (EHR, Lab, Chemistry)	- Structured, machine-readable clinical and experimental information - Enables precise reasoning and code generation - Widely adopted in biomedical informatics	- Requires schema and context understanding - Limited in capturing complex temporal or contextual relationships - Integration with unstructured data can be challenging	Risk score computation, treatment response prediction, trial eligibility reasoning
Biochemical Sequences (DNA/RNA/Protein)	- Encodes structural and functional biological information - Availability of large-scale omics datasets - Supports automated sequence analysis and engineering	- Requires specialized tokenization and domain adaptation - State-of-the-art performance often needs instruction-tuning or hybrid models - Interpretation often requires expert input	Gene variant annotation, protein engineering, chromatin accessibility analysis
Visual Data (Bioimages, Slides)	- Captures spatial and morphological context crucial for biomedical analysis - Enables agent-driven, conversational image interpretation - Supports multi-modal integration with text	- Vision encoder and LLM alignment is technically challenging - Limited availability of large, well-annotated biomedical datasets	Pathology/radiology interpretation, figure captioning, bioimage triage, question answering
Knowledge Graphs / Structured KBs	- Encodes explicit relationships among biomedical entities - Enables knowledge-grounded and explainable reasoning - Facilitates hypothesis generation	- Often incomplete or outdated - Domain-specific ontologies are hard to align and integrate - Construction and curation require substantial effort	Disease-gene association analysis, Hypothesis generation, Biomedical entity linking
Omics Datasets / Metadata Repositories	- Enables large-scale, multi-layer biological analysis - Rich source of patient- and population-level signals - Facilitates automated data curation and integration	- Requires complex preprocessing and normalization - Biological interpretation depends on expert knowledge - Data heterogeneity and quality issues are common	Multi-omics analysis, metadata curation, single-cell and population studies

Table E.1: Strengths and limitations of major data modalities in biomedical and chemical AI Agents.

Integration Strategy	Strengths	Limitations	Use Cases
Data-level Integration	- Straightforward implementation leveraging text modality - Utilizes standardized sequence and chemical formats - Fully compatible with LLM pretraining and prompting pipelines - Highly scalable with large corpora	- Discards modality-specific inductive biases - Susceptible to loss of structural information - Requires comprehensive and well-curated multimodal corpora for effective pretraining/finetuning	Encoding biological sequences and small molecules as text for LLM-based modeling, e.g., protein, molecules; TxGemma on Therapeutic Data Commons (TDC)
Contrastive Learning	- Enables robust cross-modal alignment and zero-shot transfer - Effective for retrieval and embedding-based tasks - Well-suited for aligning biomedical or chemical modalities	- Highly dependent on large, high-quality paired datasets - Limited applicability in domains with scarce aligned data	Biomedical image-text alignment, molecule-protein binding representation, radiology report generation, molecular captioning
Feature-level Fusion	- Supports modality-specific encoders (e.g., sequence, graph, vision) - Flexible, allows tailored feature extraction and late fusion - Facilitates integration of heterogeneous biomedical data	- Cross-modal interactions is weak with late fusion - Fusion strategy requires careful design and hyperparameter tuning - Reduce interpretability	Integrating protein sequence and graph features, projecting image features via vision encoders, mapping DNA features into language model space
Model-level Integration	- Modular and extensible system-level design - Enables orchestration of domain-specialized agents and external computational tools - Supports complex reasoning, planning, and workflow automation	- Overall performance bottlenecked by controller LLM - Increased system complexity and engineering overhead - Risk of error propagation across components	LLM-coordinated multi-agent frameworks for protein engineering, integrating segmentation, grounding, and coding tools for quantitative analysis

Table E.2: Strengths and limitations of multimodal integration strategies, with representative biomedical and chemical domain use cases.

Paradigm	Method	Strengths	Limitations	Use Cases
Tool Learning	API/Code Library Integration	- Direct access to domain-specific knowledge - High precision	- Integration overhead - Maintenance burden	Database querying, automated data retrieval, property prediction
	Simulator/Physical Platform	- Enables validation/testing in virtual or real environments	- Simulation accuracy/setup complexity	Virtual screening, molecular simulation, robotics-based experimentation
	Reasoning and Acting	- Flexible and extensible - Automates complex logic and tool interaction	- Error propagation - Security and reliability concerns	Automating analysis pipelines, workflow customization, code-based automation
	Retrieval-Augmented Generation (RAG)	- Reduces hallucination - Provides up-to-date and domain-specific info	- Corpus quality critical - Adds latency and complexity	Evidence-grounded reasoning, literature-aware diagnosis, knowledge augmentation
Multi-Agent Collaboration	Role-playing/Expert Simulation	- Emulates team decision making - Promotes diversity of opinion	- Role confusion - Coordination overhead	Group decision making, simulated expert discussion, virtual consultation
	Consensus/Dispute Resolution	- Increases diagnostic accuracy - Reduces individual bias	- Needs consensus protocol	Consensus-driven analysis, dispute resolution, group validation
	Adaptive/Task-based Collaboration	- Dynamically adjusts collaboration to task complexity	- Less transparent	Adaptive resource allocation, task-dependent teamwork
Self-correction & Iterative Refinement	Self-correction & Iterative Refinement	- Improves reliability/accuracy - Enables continual improvement	- Increases computation cost - Needs feedback mechanisms	Error correction, iterative plan refinement, automatic code/debug improvement

Table E.3: Overview of major paradigms and representative methods for agentic AI in biomedical and chemical domains, highlighting their strengths, limitations, and typical use cases.

Tuning Strategy	Strengths	Limitations	Use Cases
Instruction Tuning (SFT)	- Infuses domain-specific knowledge via task-relevant instructions - Enhances instruction-following and tool selection capabilities - Straightforward to implement for various biomedical and chemical tasks	- Requires high-quality labeled instruction–output pairs - Limited generalization to unseen tasks or new tools	Multi-tool and multi-modal instruction tuning for medical agents Specialized chemical or therapeutic models finetuned with domain-specific prompts
Reinforcement / Preference-based Tuning (RL, DPO)	- Enables interaction with external environments for autonomous policy learning - Supports optimization for multi-objective or delayed rewards - Preference-based tuning (e.g., DPO) allows alignment with expert or human feedback without explicit reward design	- RL is sensitive to reward function design and can be computationally expensive - Exhibits instability in complex or high-dimensional tasks - Preference-based tuning requires curated pairwise preference data and has limited interpretability	RL: Chemical generation and property optimization, environment interaction Preference-based: Energy/structure optimization via DPO or related methods

Table E.4: Strengths and limitations for domain-specific tuning strategies in biomedical and chemical AI agents.

Application Area	Representative Task Types	Typical Input Modalities
Scientific Discovery & Experiment Automation	- Hypothesis generation - Molecular/protein design - Gene editing setup - Variant analysis - Bioinformatics automation - Drug/material design - Property prediction	- Molecular graphs - Biological sequences - Omics tables - Natural language instructions - Virtual environment - Robotic tools
Clinical Decision Support & Reasoning	- Diagnosis assistance - ICU risk scoring - Emergency triage - Pathology interpretation - Multilingual doctor-patient communication	- Clinical notes (EHR) - Structured patient data - Medical imaging - Guidelines - User instructions (multi-language) - Structured checklists
Biomedical and Chemical Question Answering (QA) and Data Analysis	- Answering clinical/Chemical queries - Multi-hop retrieval - Justified reasoning grounded in medical knowledge - Information Extraction - Visual analytics coordination	- Medical/Chemical questions - Domain literature - Knowledge bases - Biomedical APIs - Bioimage files - Calculator inputs - Interface instructions

Table E.5: Task types and input modalities for different application areas.

Section	Type	What is Measured	Typical Metrics	Strengths	Limitations	Examples
Core Knowledge Reasoning	QA Benchmarks	Domain knowledge reasoning	Accuracy, F1, ROC-AUC	Quantitative, objective	Limited to static knowledge	MedQA, MedMCQA, PubMedQA, MMLU
	Classification / Regression Benchmarks	Property prediction	Accuracy, Precision, Recall, ROC-AUC	Standardized tasks	Not reflective of real workflows	MoleculeNet, LAB-Bench
Task-Oriented Evaluation	Task-specific Benchmarks	End-to-end workflow / task completion	Task-specific accuracy, Hit Rate, Human evaluation	Captures real-world process	Benchmark construction is expensive	BioDiscoveryAgent, Biomni, BioDSA-1K
	Interactive Benchmarks	Agent interaction, adaptation, process improvement	Diagnostic accuracy over time, Entity overlap	Measures dynamic behavior	Scenario design impacts result	Agent Hospital, AI Hospital
	Real-world Evaluation (Wet-lab)	Real-world task success, lab experiment outcomes	Yield, Purity, Task success rate, Expert assessment	Highest ecological validity	Expensive, low throughput	Biomni (wet-lab), Coscientist, ChemCrow

Table E.6: Summary of objective evaluation strategies, measured abilities, metrics, and representative benchmarks for biomedical and chemical AI agents.

Method	How it Works	Strengths	Limitations	Examples
Human/LLMs Ranking	- Interactive arenas with head-to-head ranking of agent outputs by clinicians or LLMs; - Uses pairwise comparisons and statistical ranking models (e.g., Bradley–Terry)	- Direct comparative assessment - Captures user/clinician preferences - Scalable via LLM annotators	- Expensive and slow for human annotation - LLM judges introduce bias	MedArena, Decentralized Arena, Chatbot Arena
Human Qualitative Evaluation	- Experts (often blinded) qualitatively review - Score agent outputs using rubrics or structured protocols	- Gold standard for complex and open-ended tasks - Assesses utility and interpretability	- Not scalable - Resource-intensive	Biomni (expert review), ChemCrow, GeneAgent
LLM-as-Judge	- LLMs (e.g., GPT-4, Claude) are prompted to review and score outputs using detailed rubrics or Likert scales - Assess clarity, relevance, novelty, etc.	- Scalable & Systematic - High agreement with expert humans	- Dependent on rubric/prompt quality - Possible model bias	AI Hospital, ResearchAgent, LLM4BioHypoGen, Bixbench

Table E.7: Comparison of mainstream subjective evaluation strategies for agentic AI systems in biomedical and chemical domains, summarizing workflows, strengths, limitations, and representative systems.

Agent Name	Type	Modality	Agent Ability
ProtAgents (Ghafarollahi and Buehler, 2024)	Multi-agent	Protein structure (3D), Sequence, Text	Interacts with simulated protein environments through natural language by selecting sequences, proposing mutations, and optimizing structural outcomes
A-Lab (Szymanski et al., 2023)	Single-agent	Text + Robotic control + XRD signal	Autonomously proposes, executes, and optimizes the synthesis of materials using robotics and active learning
Chemistry42 (Ivanenkov et al., 2023)	Single-agent	Molecular (2D, 3D), structural data	Autonomously generates de novo drug-like molecules with desired properties using 30+ generative models, reinforcement learning, and medicinal chemistry filters
Eunomia (Ansari and Moosavi, 2024)	Single-agent	Text, Tools	Extracts structured materials datasets (e.g., doping relationships, MOF formulas, water stability) from unstructured scientific text
CACTUS (McNaughton et al., 2024)	Single-agent	Text, Tools	Integrates cheminformatics tools with LLMs via LangChain to answer chemistry questions, estimate molecular descriptors, and assist drug discovery through zero-shot reasoning and tool selection
ACEGEN (Bou et al., 2024)	Single-agent	Text, SMILES, SELFIES	Generates and optimizes drug-like molecules using reinforcement learning with customizable scoring functions
ChemLangAgent (Boiko et al., 2023)	Multi-agent	Text, SMILES	Integrates an LLM with tools to plan reactions, select reagents, analyze properties, and assess safety
ChemCrow (M. Bran et al., 2024)	Multi-agent	Text, SMILES, Tools	Calls 18 chemistry tools to perform reagent lookup, synthesis planning, property retrieval, safety assessment, and explain reasoning in natural language
ChatMOF (Kang and Kim, 2024)	Multi-agent	Text, Tools	Predicts properties, searches database info, and generates MOF structures with target properties using LLM-coordinated tools
LLM-RDF (Ruan et al., 2024)	Multi-agent	Text, Code, Instrument control	Automates synthesis tasks like literature mining, experiment design, execution, and result interpretation
Deep Thought (Smbatyan et al., 2025)	Multi-agent	Structured data, Code, and Text	Autonomously performs end-to-end virtual screening: plans strategy, writes and executes code, selects molecules to query, and submits results iteratively
AtomAgents (Ghafarollahi and Buehler, 2025)	Multi-agent	Text, Code, Image, Simulation	Automates alloy design by integrating multimodal data, running atomistic simulations, analyzing plots, retrieving knowledge, and validating hypotheses
X-LoRA-Gemma (Stewart and Buehler, 2025)	Multi-agent	Text, Molecular Structures, Scientific Data	Analyzes, designs, and validates molecules with desired properties through human–AI and AI–AI interaction, performs inverse design by tuning properties such as dipole moment and polarizability, and generates candidate molecular structures using generative modeling
MT-MOL (Kim et al., 2025)	Multi-agent	Text, Tools	Designs molecules using tool-guided, stepwise reasoning across four roles: tool selection, molecule generation, consistency verification, and structured review feedback
DrugAgent (Liu et al., 2024)	Multi-agent	Text, Structured data	Autonomously identifies domain-specific requirements, builds reusable tools, explores and refines multiple modeling strategies, performs code execution and debugging, and ultimately selects the most effective solution
TxGemma / Agentic-Tx (Wang et al., 2025a)	Agentic-Tx: Multi-agent TxGemma: Single-agent	Structured (SMILES, protein/nucleotide sequences), Text	TxGemma predicts therapeutic properties across 66 tasks, including toxicity, ADME, drug synergy, and AE prediction; enables reasoning via natural language with scientific explanations Agentic-Tx orchestrates complex workflows using 18 tools for property prediction, literature retrieval, molecule analysis, and trial planning.
PharmAgents (Gao et al., 2025)	Multi-agent	Text, Molecular Structure	Simulates the full drug discovery pipeline using LLM-driven agents integrated with ML tools, performing disease-target mapping, compound generation/optimization, and in silico preclinical assessment with explainable and evolving outputs.
ResearchAgent (Baek et al., 2025)	Multi-agent	Text, Tools	Automatically generates, evaluates, and iteratively refines research ideas (problem, method, experiment) using literature, entity knowledge, and multi-agent reviewing feedback

Table E.8: Overview of representative AI agents applied in chemical research, highlighting their design paradigms, core functionalities, and integration of language models with domain-specific tools and data.

Agent Name	Type	Modality	Agent Ability
ArabicAgent (ALMutairi et al., 2024)	Single-agent	Multilingual text	Generates plausible biomedical hypotheses from background knowledge and research questions and is evaluated for novelty, validity, and relevance
AutoBA (Zhou et al., 2024a)	Single-agent	Text, Structured omics data	Enables end-to-end multi-omics data analysis by autonomously planning, coding, executing, and debugging bioinformatics workflows with minimal user input.
Omega (Royer, 2024)	Single-agent	Text, Vision	Interactively analyzes bioimages; segments cell nuclei, counts and measures objects, debugs and edits code, creates custom widgets, and executes follow-up analyses; supports visual inspection, tool invocation, and on-demand UI / code generation through conversational input
AutoPE (Liu et al., 2025)	Single-agent	Protein sequence and Protein structure	Automates multimodal AutoML (model selection, hyperparameter tuning, data retrieval) for protein engineering via natural language input
GeneAgent (Wang et al., 2024g)	Single-agent	Text, Databases	Generates and verifies biological process names for gene sets by autonomously interacting with 18 domain-specific biomedical databases, reducing hallucinations and improving reliability
FG-RAG (Selinger et al., 2024)	Single-agent	Text, Omics, Chemical structure	Autonomously plans, executes, and interprets focal graph searches using LLMs
GeneGPT (Jin et al., 2024b)	Single-agent	Text, Tools	Executes real-time Web API calls to NCBI tools to answer genomics-related queries, including multi-hop reasoning via chain-of-thought API sequences
CRISPR-GPT (Huang et al., 2024)	Single-agent	Text, Tools	Supports researchers in CRISPR experiment planning by automating key steps like system selection, guide RNA design, and validation setup, using modular interactive workflows
BIA (Xin et al., 2024)	Single-agent	Text, RNA sequences	Executes end-to-end single-cell RNA-seq pipelines: retrieves datasets, extracts structured metadata, generates bioinformatics workflows, adapts code, and reports results autonomously
Biomni (Huang et al., 2025a)	Single-agent	Structured data, Text, Code, Multi-omics	Executes biomedical tasks autonomously, including gene prioritization, variant analysis, protocol generation, multi-omics integration, and hypothesis generation from real-world data
ChatNT (de Almeida et al., 2025)	Single-agent	Text, DNA sequences	Performs English-based classification and regression across 27 biological sequence tasks (DNA, RNA, protein); supports multitask inference and interpretable predictions via a conversational interface
ESCARGOT (Matsumoto et al., 2025)	Single-agent	Text, Structured knowledge graphs	Dynamically generates and executes reasoning strategies over biomedical knowledge graphs using Graph-of-Thoughts and Cypher queries to improve factuality and reduce hallucinations
MedMax (Bansal et al., 2025)	Single-agent	Text + Image	Understands and generates interleaved biomedical image-text content; performs multimodal VQA, generates medical reports, images, and answers visual queries across diverse biomedical tasks
BioDiscoveryAgent (Roohani et al., 2024)	Multi-agent	Text, Tabular data	Assists biomedical research by answering complex questions through literature retrieval, table analysis, and reasoning
BioImage (Lei et al., 2024)	Multi-agent	Text, Code, Image	Navigates bioimaging resources, answers technical questions, generates and executes code, performs image segmentation, queries databases, and analyzes results autonomously using tool-calling and vision
LLM4BioHypoGen (ALMutairi et al., 2024)	Multi-agent	Text, Tools	Generates novel biomedical hypotheses from literature by multi-agent collaboration and tool learning
Virtual Lab (Swanson et al., 2024)	Multi-agent	Text, Tools	Designs nanobody candidates for SARS-CoV-2 variants via interdisciplinary collaboration, performs reasoning, code writing, and experimental planning for protein engineering

Table E.9: Part 1: Overview of AI agents in biomedical, medical, and clinical domains, highlighting their application areas, interaction modalities, and agent design strategies.

Agent Name	Type	Modality	Agent Ability
RankNorm (Fan et al., 2025a)	Multi-agent	Text, Database	Performs terminology normalization by retrieving and ranking candidate terms using multi-agent LLM collaboration, addressing ambiguity in short social media texts without training
Intelliscope Agent System (Aamer et al., 2025)	Multi-agent	Text, Knowledge graphs	Explores biomedical knowledge graphs using semantic search, designs AI predictors via multi-agent deliberation (Analyst, Scientist, Reviewer), and refines research proposals iteratively
scBaseCount (Youngblut et al., 2025)	Multi-agent	Text, Genomics	Constructs a large-scale single-cell transcriptomics data repository for modeling by automated data mining, metadata extraction, and uniform gene expression processing
AI co-scientist (Gottweis et al., 2025)	Multi-agent	Text, Tools	Generates, critiques, evolves, and ranks novel scientific hypotheses; collaborates with scientists to propose research plans, especially for drug repurposing, target discovery, and mechanistic explanations
ROBIN (Ghareeb et al., 2025)	Multi-agent	Text, Tools	Multi-step biomedical reasoning agent that plans, retrieves evidence, reasons, and integrates results for research insights.
EHRAgent (Shi et al., 2024)	Single-agent	Structured tabular, Text	Translates clinical questions into executable code for multi-table EHR reasoning with minimal supervision
PathChat (Lu et al., 2024b)	Single-agent	Image, Text	Answers pathology-related questions using histology images and natural language, suggests differential diagnoses and IHC tests, interprets image morphology, and supports interactive multi-turn dialogue for clinical decision-making and education.
ClinicalRAG (Lu et al., 2024c)	Single-agent	Text, Tools, Knowledge graphs	Enhances clinical diagnosis by extracting medical entities, retrieving heterogeneous knowledge, generating natural language summaries, and integrating references to support LLM reasoning.
AgentMD (Jin et al., 2024a)	Single-agent	Text, Code	Automatically curates clinical calculators from literature, selects and applies appropriate tools to patient notes for risk estimation, and answers risk-related clinical questions
i-MedRAG (Xiong et al., 2024b)	Single-agent	Text, Tools	Iteratively generates and answers follow-up queries to enhance LLM reasoning in complex medical QA scenarios, improving accuracy beyond vanilla RAG methods.
MedLingua (Himel et al., 2024)	Single-agent	Text, Speech	Understands symptoms in multiple languages, provides treatment and doctor recommendations, resolves ambiguities in similar-sounding terms, supports automatic translation, and interacts via text or voice.
MeNTi (Zhu et al., 2025)	Single-agent	Text, Tools	Selects appropriate medical calculators, fills parameters, performs unit conversions, and computes results via nested tool calling using a specialized medical toolkit
MMedAgent (Li et al., 2024a)	Multi-agent	Text, Image	Interprets medical images and texts, retrieves domain knowledge, performs diagnosis and treatment reasoning, and solves complex medical QA tasks with tool assistance.
Agent hospital (Li et al., 2024b)	Multi-agent	Text, Knowledge bases, Tools	Simulates roles of healthcare workers to collaboratively diagnose and treat virtual patients via multi-round dialogue and tool use
Medical Necessity Justification (Pandey et al., 2024)	Multi-agent	Text, Structured Data	Collaboratively analyze patient records, identify supporting evidence, and justify medical procedures using agent-role coordination
MDAgents (Kim et al., 2024)	Multi-agent	Text, Image, Video	Dynamically forms LLM teams (solo or collaborative) based on medical query complexity; performs complexity classification, specialist recruitment, collaborative discussion, and final decision-making.
AI Hospital (Fan et al., 2025b)	Multi-agent	Text, Structured Data	Simulates multi-turn clinical diagnosis using LLM-powered Doctor agents interacting with Patient and Examiner agents to collect symptoms, recommend tests, and generate diagnostic reports

Table E.10: Part 2: Overview of AI agents in biomedical, medical, and clinical domains, highlighting their application areas, interaction modalities, and agent design strategies.

Agent Name	Type	Modality	Agent Ability
MedAgent-Pro (Wang et al., 2025d)	Multi-agent	Text, Medical Image	Implements a structured, evidence-based workflow for medical diagnosis, combining guideline-driven planning with patient-specific step-by-step reasoning using RAG, segmentation tools, coding modules, and VLMs
MedRAX (Fallahpour et al., 2025)	Single-agent	Text, Image	Dynamically integrates specialized tools to perform step-wise reasoning and tool-based decision making for complex chest X-ray interpretation, including disease detection, localization, diagnosis, and report generation
Clinicalagent (Yue et al., 2024)	Multi-agent	Text, Structured data	Decomposes complex clinical trial queries into subproblems and solves them using specialized agents and external tools to generate accurate and explainable results
TRIAGEAGENT (Lu et al., 2024a)	Multi-agent	Text, Databases	Enables zero-shot triage through role-playing agents using retrieval-augmented generation, multi-stage collaboration, and confidence-based consensus
ColaCare (Wang et al., 2025c)	Multi-agent	Structured data, Text	Simulates multidisciplinary clinical consultations by combining expert EHR models and LLM agents to generate interpretable, personalized clinical predictions
Multi-Agent Conversation (Chen et al., 2025b)	Multi-agent	Text, Databases	Simulates a team of doctor agents and a supervisor agent to collaboratively reason through complex, rare disease cases, improving diagnostic accuracy and test recommendations over single-agent LLMs

Table E.11: Part 3: Overview of AI agents in biomedical, medical, and clinical domains, highlighting their application areas, interaction modalities, and agent design strategies.

Survey	Domain	Scientific Aim
Gao et al. (2024)	Medical	Proposes AI agents that act as AI scientists by integrating large language models, machine learning tools, and experimental platforms to support and automate biomedical research
Visan and Negut (2024)	Medical/Chemical	Reviews the multifaceted role of AI in accelerating and optimizing the drug discovery and delivery process, including AI applications in target identification, virtual screening, drug design, property prediction, drug repurposing, and combination therapy
Tom et al. (2024)	Chemical	Reviews the emerging field of self-driving laboratories that integrate automation, machine learning, and closed-loop experimentation for accelerating discovery and optimization in chemistry and materials science
Ren et al. (2025)	Multi-domain (across biomedical, chemical, clinical, etc.)	Summarizes the design, architecture, applications, benchmarks, and ethical considerations of LLM-based scientific agents designed to automate complex research tasks, including hypothesis generation, experimental design, data analysis, and simulation
Ramos et al. (2025)	Chemical	Summarizes the use of LLMs and autonomous agents in accelerating molecule design, property prediction, synthesis planning, and automation in chemistry, and discusses their architectures, capabilities, challenges, and broader scientific applications.
Schouten et al. (2025)	Medical	Analyzes 432 papers (2018–2024) on deep learning-based multimodal AI in medicine, summarizing its development, clinical applications, fusion methods, and key challenges, aiming to guide its integration into clinical practice
Zheng et al. (2025)	Multi-domain (across biomedical, chemical, clinical, etc.)	Categorizes the evolving roles of LLMs in scientific discovery through a three-level autonomy framework (Tool, Analyst, Scientist) and maps their applications to the six stages of the scientific method, highlighting the shift from automation tools to autonomous research agents

Table E.12: Summary of recent survey papers on AI agents in research, outlining their domains of focus and core objectives in advancing automation and autonomy across disciplines.

Dataset	Domain	Modality	Size	Task
BioTrove (Yang et al., 2024a)	Biology	Image + Text	161.9 million images	Fine-grained image classification
VLM4Bio (Maruf et al., 2024)	Biology	Image + Text	∼30K images, ∼469K QA pairs	Species classification, trait identification, trait grounding, trait referring, trait counting, VQA, multimodal reasoning
LLM4BioHypoGen (ALMutairi et al., 2024)	Clinical	Text	207 dialogue-note pairs + synthetic Arabic dialogues	Medical dialogue generation from clinical notes; Training/evaluating Arabic NLP models
MIRAGE (Xiong et al., 2024a)	Medical	Text	7,663 QA examples from 5 datasets	Multiple-choice medical question answering (QA)
CliMedBench (Ouyang et al., 2024)	Clinical	Text	33,735 QA instances	clinical QA, reasoning, summarization, knowledge application, info retrieval, hallucination detection, toxicity detection
BIOMEDICA (Lozano et al., 2025)	Biomedical / Medical / Clinical / Biology	Image + Text	24,076,288 image-caption pairs from 6,042,494 articles	Image classification, retrieval (image-to-text, text-to-image), zero-shot classification, pretraining for VLMs
MicroVQA (Burgess et al., 2025)	Biomedical	Image + Text	1,042 VQA samples	Multimodal visual question answering (VQA)
MedAgentGym (Xu et al., 2025)	Biomedical	Tabular, Text, Code, Sequences	72,413 instances across 129 categories	Code generation for medical reasoning
ReasonMed (Sun et al., 2025)	Medical	Text	370K high-quality samples	Medical question answering with complex reasoning
ChemBench (Mirza et al., 2025)	Chemical	Text	∼2,788 QA pairs	QA; human preference judgment
BixBench (Mitchener et al., 2025)	Biomedical	Text, Code, Tabular data	53 analytical capsules, 296 open-ended questions	Complex analytical reasoning, scientific data analysis, multi-step problem-solving, agentic code execution, QA
BioDSA-1K (Wang et al., 2025b)	Biomedical	Structured data	1,029 hypotheses, 1,177 tasks from 328 publications	Hypothesis validation using code generation, reasoning, statistical analysis, and data interpretation

Table E.13: Summary of recent datasets on AI agents in biomedicine and chemistry domains.