Briefings in Bioinformatics, 2026, 27, bbag110
https://doi.org/10.1093/bib/bbag110
Published: 25 March 2026
Review

OXFORD

Large language model agents for biological intelligence across genomics, proteomics, spatial biology, and biomedicine

Sajib Acharjee Dip[1,*], Dipanwita Mallick[2], Uddip Acharjee Shuvo[3], Shovito Barua Soumma[4,5], Fazle Rafsani[4], Bikash Kumar Paul[1], Nazifa Ahmed Moumi[1], Shafayat Ahmed[1], Liqing Zhang[1,6,7,*]

[1]Department of Computer Science, Virginia Tech, 225 Stanger Street, Blacksburg, VA 24061, United States
[2]Department of Computer Science, North South University, Bashundhara, Dhaka 1229, Bangladesh
[3]Institute of Information Technology, University of Dhaka, Nilkhet Road, Dhaka 1000, Bangladesh
[4]School of Computing and Augmented Intelligence, Arizona State University, 699 S Mill Ave, Tempe, AZ 85281, United States
[5]College of Health Solutions, Arizona State University, 550 N 3rd Street, Phoenix, AZ 85004, United States
[6]Fralin Biomedical Research Institute at VTC, Virginia Tech, 2 Riverside Circle, Roanoke, VA 24016, United States
[7]Fralin Biomedical Research Institute Cancer Research Center, Virginia Tech, Washington, DC 20005, United States

*Corresponding authors. Sajib Acharjee Dip, Department of Computer Science, Virginia Tech, 225 Stanger Street, Blacksburg, VA 24061, United States. E-mail: sajibacharjeedip@vt.edu; Liqing Zhang, Department of Computer Science, Virginia Tech, 225 Stanger Street, Blacksburg, VA 24061, United States. E-mail: lqzhang@cs.vt.edu.

Abstract

Large language models (LLMs) are evolving from passive predictors into agentic systems capable of planning, tool-use, and multimodal reasoning. This shift is especially consequential for biology, where complex, noisy, and multi-scale data require adaptive and integrative computational strategies. In this review, we provide the first systematic synthesis of LLM-based agents across genomics, molecular biology, imaging, biomedical analysis, and automated bioinformatics workflows. We analyze >60 emerging systems and organize them within a unifying framework that characterizes agentic traits, such as autonomous decision-making, external tool invocation, memory, and self-correction. Across domains, agentic LLMs show early promise in enabling multi-step analysis, linking heterogeneous evidence, and supporting exploratory scientific tasks. At the same time, our comparative assessment highlights consistent challenges, including unstable reasoning, limited biological grounding, retrieval misalignment, and barriers to reproducibility and biosafety. We conclude by outlining opportunities for trustworthy and collaborative biological agents, including multimodal integration, closed-loop experimental design, and robust evaluation practices. This survey aims to clarify the emerging landscape and chart a path toward reliable agentic systems for biological discovery.

Keywords **large language models, agentic AI, bioinformatics, genomics, proteomics, drug discovery, multi-omics, metagenomics, LLM agents, biological sequence analysis, transformer models, biomedical AI, artificial intelligence in biology**

Introduction

Biological research is undergoing a rapid transition from static, single-step computational analyses toward complex, multi-stage workflows that integrate genomics, proteomics, structural modeling, drug design, spatial biology, and clinical imaging. Modern biological questions increasingly demand the orchestration of heterogeneous tools, databases, and analytical procedures: aligning sequences, running statistical models, retrieving literature, generating molecular candidates, executing simulations, and iteratively refining hypotheses. These workflows resemble scientific "procedures" rather than isolated prediction problems, requiring planning, decision-making, and continuous integration of new evidence.

Large language models (LLMs) have begun to transform this landscape. Beyond text generation, contemporary LLMs can write and debug code, call external tools, query knowledge bases, reason over intermediate results, and provide interpretable narratives of complex analyses. When embedded in control loops that allow them to plan actions, invoke software, self-correct errors, and adapt to new information, these models function as *agentic* systems capable of multi-step scientific reasoning and autonomous workflow execution.

Over just the past 2 years, a new class of *agentic biological systems* has emerged across domains including drug discovery, molecular design, gene expression analysis, single-cell and spatial biology, structural prediction, and medical imaging. These systems differ fundamentally from conventional LLM applications: rather than responding to isolated prompts, they coordinate multi-tool pipelines, maintain internal state, and perform iterative decision-making. This shift raises new opportunities including greater automation, reproducibility,

Received: December 3, 2025. **Revised:** January 10, 2026. **Accepted:** February 10, 2026

Table 1 Positioning of this review relative to representative recent surveys on LLMs, agents, and scientific discovery.

Survey	Agent-centric	Cross-domain biology	Evaluation operationalization	Failure/risk synthesis
Ruan *et al*. (2026) [28]	–	Partial	Limited	Limited
Yang *et al*. (2025) [29]	✓	Partial	Limited	–
Zhou *et al*. (2025) [32]	–	Partial	–	–
Guo *et al*. (2025) [1]	–	–	Limited	Limited
Zhang *et al*. (2024) [9]	–	Partial	Limited	–
Chen *et al*. (2024) [30]	✓	–	Limited	–
Wei *et al*. (2024) [31]	✓	–	–	–
LLMAgent4Bio	✓	✓	✓	✓

interpretability but also introduces challenges in grounding, verification, safety, and integration with experimental workflows.

Despite the pace of development, the literature remains fragmented. Existing surveys typically address (i) foundation models for biology [1–9], (ii) modality-specific applications such as radiology [10–17] or genomics [9, 18–23], or (iii) general-purpose AI agents [24–27]. However, none provide a unified treatment of *agentic* behavior in biological analysis: how systems plan, act, use tools, collaborate, retrieve evidence, verify outputs, or integrate multimodal biological knowledge. In particular, existing surveys in biology and biomedicine largely treat LLMs as static predictors or multimodal encoders, with limited discussion of system-level behaviors that arise when models are embedded in multi-step analytical workflows [2, 5, 6, 28]. Surveys of foundation and scientific LLMs emphasize pretraining strategies, representation learning, and downstream task performance across biological and chemical domains [8, 9], while domain- or modality-specific reviews focus on application-level gains within narrower settings such as radiology, genomics, or drug research and development [1, 10, 18]. Recent surveys on autonomous or agentic systems instead prioritize abstract agent architectures, planning mechanisms, and software frameworks for automation [29–31]. While these works provide important conceptual foundations, they remain largely domain-agnostic and do not address biology-specific constraints such as biological grounding, experimental uncertainty, ontology alignment, tool reliability, or safety-critical failure modes. As a result, critical questions surrounding agent coordination, verification, error propagation, and the evaluation of multi-step biological reasoning workflows remain largely unaddressed in prior reviews. This survey fills this gap by treating agentic behavior itself as the primary unit of analysis and by providing a unified, cross-domain synthesis of biological LLM agents grounded in biological structure, evaluation practice, and failure modes (Table 1). As biological AI rapidly shifts from passive models to active computational collaborators, a cross-domain synthesis is urgently needed.

This survey fills that gap. Unlike prior surveys, our analysis treats agentic behavior itself as the primary unit of comparison, rather than model architecture or task performance in isolation. Our contributions are four-fold:

- **We introduce the first cross-domain conceptual framework for *agentic* LLM systems in biology**, distinguishing them from foundation models, tool-augmented LLMs, and domain-specific multimodal architectures.
- **We analyze 60+ recent systems across genomics, proteomics, molecular modeling, spatial biology, drug discovery, and clinical imaging**, providing the most comprehensive synthesis to date of how agentic behaviors are implemented and evaluated.
- **We propose a taxonomy of biological agent systems** organized along task-type, agent-role, architectural design, and domain grounding, enabling structured comparison across highly heterogeneous systems.
- **We identify evaluation gaps, risks, and open challenges** through the first cross-domain synthesis of evaluation practices and failure modes for biological LLM agents, highlighting limitations in tool reliability, hallucination control, biological grounding, and the absence of standardized benchmarks for multi-step agentic workflows.

Together, these contributions articulate a unified perspective on how LLM agents are reshaping computational biology and scientific automation. Rather than duplicating domain-specific details presented later in the paper, the introduction highlights the conceptual forces driving the transition from foundation models to agentic biological systems and outlines the opportunities for automated, interpretable, and reproducible scientific discovery enabled by this emerging paradigm. Unlike prior LLM or modality-specific surveys, we introduce the first unified taxonomy of agent roles, architectural patterns, and biological domain grounding, supported by the first cross-domain evaluation matrix and failure-mode analysis. A high-level overview of the agentic LLM ecosystem covered in this review is illustrated in Fig. 1, including task categories, architectures, biological domains, evaluation criteria, and representative systems.

Large language models and agentic paradigms in biology

LLMs have moved rapidly from passive text generators to systems capable of planning, tool-use, and multimodal reasoning. Unlike earlier biomedical foundation models such as BioGPT, ProteinGPT [33], or ProtLLM, which primarily provided descriptive answers, emerging biological agents operate as controllers that coordinate software tools, integrate heterogeneous evidence, and adaptively refine analyses.

From foundation models to large language model agents

Traditional models learned statistical patterns from sequences or text but were limited in tasks requiring computation or multi-step workflows. Agentic systems address this by embedding an LLM inside a control loop that enables:

1. task decomposition and planning;
2. execution of external tools (e.g. aligners, structure predictors, and spatial-omics models);

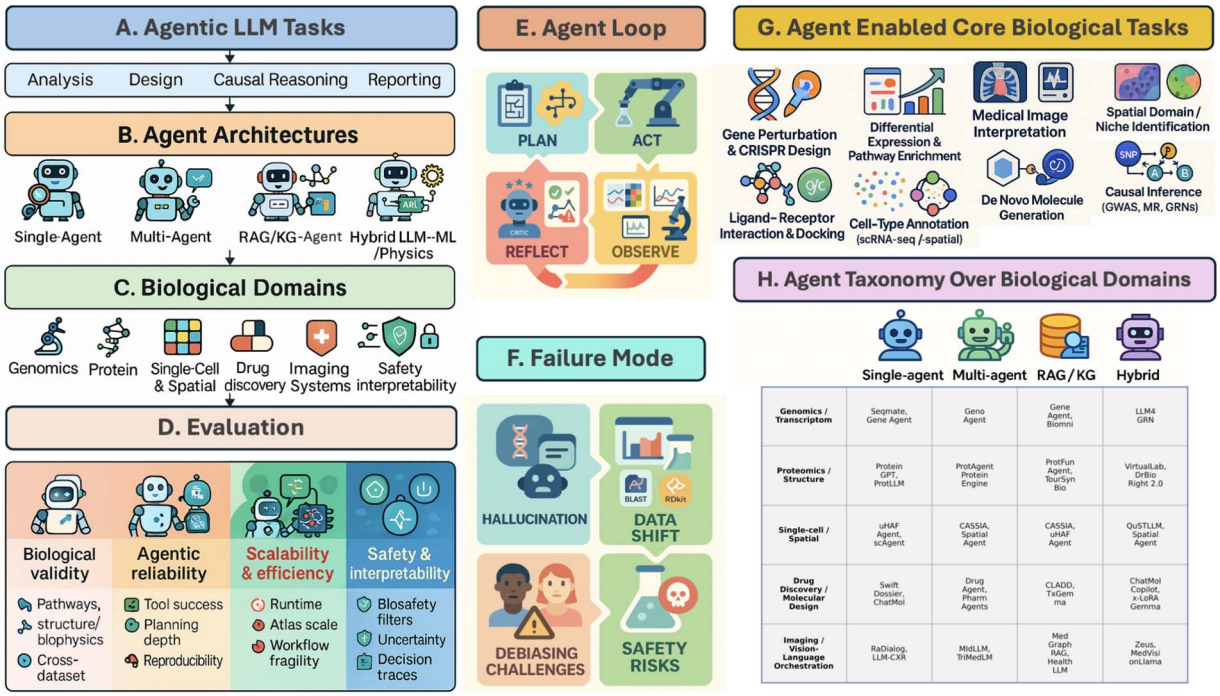

Figure 1 Overview of the LLMAgent4Bio conceptual framework. The figure summarizes core components of biological LLM-agent systems, including: (A) major task categories (analysis, design, causal reasoning, and reporting); (B) dominant agent architectures (single-agent, multi-agent, retrieval-centric, and hybrid LLM–ML/physics systems); (C) biological application domains; (D) evaluation dimensions spanning biological validity, agentic reliability, scalability, and interpretability; (E) a generic PLAN–ACT–OBSERVE–REFLECT agent loop; (F) common failure modes encountered in biological agents; (G) core biological tasks performed by agentic LLM systems; and (H) a taxonomy of representative LLM-agent systems across genomics, proteomics, single-cell/spatial biology, drug discovery, and imaging workflows.

3. interpretation of intermediate results;
4. iterative refinement and error recovery.

This paradigm now supports diverse workflows including RNA-seq processing, gene-set interpretation, protein and structure analysis, spatial-omics annotation, and drug-design loops where the LLM provides high-level reasoning and domain tools provide precision.

Defining agentic systems in bioinformatics

Across surveyed systems, agentic biological LLMs share four minimal components:

- **Task specification**: the biological goal provided by the user or inferred by the agent.
- **External tools**: established computational software, databases, and simulators.
- **Iterative control loop**: planning, tool invocation, monitoring, and self-correction.
- **LLM backbone**: responsible for reasoning, decomposition, and interpretation.

Agentic systems span a continuum:

- **Single-agent controllers** for structured pipelines such as RNA-seq or gene-set analysis.
- **Multi-agent ecosystems** for open-ended reasoning (planner–critic–specialist roles).
- **RAG/knowledge-centric systems** for ontology-driven or retrieval-heavy tasks.
- **Hybrid LLM–ML/physics systems** that integrate AlphaFold, docking engines, or spatial deconvolution models.

Core capabilities enabling biological tasks

Despite broad domain differences, effective agents rely on five capabilities:

1. **Tool-use**: orchestrating established bioinformatics and modeling tools.
2. **Multi-step planning**: managing branching workflows common in omics and drug discovery.
3. **Memory and self-reflection**: tracking intermediate outputs and correcting errors.
4. **Knowledge grounding**: using ontologies, curated databases, and retrieval for biological plausibility.
5. **Multimodal integration**: combining sequences, structures, expression matrices, spatial coordinates, or images.

These capabilities motivate the unified taxonomy developed in Section Conceptual framework and taxonomy of agentic large language model systems that links task type, agent architecture, and biological domain grounding.

Conceptual framework and taxonomy of agentic large language model systems

Biological LLM agents are emerging across genomics, proteomics, spatial transcriptomics, drug design, and clinical reasoning, yet these systems vary widely in the biology they target, the workflows they automate, and the degree of autonomy they require. To provide conceptual clarity, we introduce a unified taxonomy that organizes biological agents along two interacting dimensions: *task type* and *agent role*, and then connect these dimensions to the

architectural patterns and biological domains in which agents operate.

Classification by task type and agent role

Task-type axis.

Across domains, agentic systems repeatedly serve four functional categories:

- **Analysis:** Processing and interpreting biological data QC, clustering, annotation, differential expression, enrichment testing, and imaging analysis. This category dominates genomics, single-cell, and spatial workflows.
- **Design and engineering:** Constructing or optimizing biological entities, such as sgRNAs, perturbation panels, protein variants, ligands, or nanobodies. These tasks require constraint handling, tool coordination, and iterative refinement.
- **Causal and mechanistic reasoning:** Inferring regulatory relationships, generating hypotheses, or integrating multimodal evidence to propose mechanistic explanations.
- **Summarization and reporting:** Producing structured experimental summaries, documentation, or narrative interpretations for downstream use.

These four categories recur consistently and define the operational logic that an agent must support. While systems sometimes span multiple categories, characterizing them by their *primary objective* clarifies distinct computational and reasoning demands.

Agent-role axis.

Biological workflows naturally decompose into specialized responsibilities. Across surveyed systems, five recurrent roles appear:

- **Planner/Coordinator:** Decomposes tasks, selects methods, and maintains state.
- **Tool Executor/Coder:** Generates code, invokes external software, and resolves errors.
- **Critic/Verifier:** Checks correctness, enforces constraints (e.g. ontology and sequence rules), and mitigates hallucinations.
- **Domain Specialist:** Applies contextual knowledge (e.g. proteomics, spatial biology, and medicinal chemistry) to interpret intermediate results.
- **Reporter/Explainer:** Produces readable summaries, interpretations, or structured output.

These roles may be implemented in a single controller (lightweight systems) or distributed among multiple LLMs (multi-agent ecosystems). Task types strongly shape role configurations; e.g. design tasks frequently require planner–critic–specialist triads, whereas analysis pipelines rely more on planner–executor pairs.

Intersection of task type and agent role.

Combining both axes yields a structured view of agentic behavior. Analytical agents prioritize robust tool execution; engineering agents emphasize iterative planning and constraint verification; and causal reasoning agents rely on literature grounding and critic modules. This intersection provides the conceptual basis for the taxonomy introduced below.

Taxonomy and visual model

Building on the two axes above, we propose a three-layer taxonomy that situates biological LLM agents within a coherent ecosystem: *task type*, *agent architecture*, and *biological domain grounding*.

Layer 1: Task type.

The four functional categorie-s: analysis, engineering, causal reasoning, and reporting form the top of the hierarchy.

Layer 2: Agent architecture.

These task types map naturally onto specific architectural patterns:

- **Single-agent controllers** for predictable pipelines (e.g. RNA-seq processing and single-cell QC).
- **Multi-agent ecosystems** for exploratory or open-ended tasks (e.g. hypothesis generation, spatial interpretation, and drug design).
- **RAG-centric/ontology-centric agents** for tasks requiring heavy grounding, such as gene-set analysis or clinical reasoning.
- **Hybrid LLM–ML/physics systems** integrating structure prediction, docking, simulations, or GRN inference.

Layer 3: Biological domain grounding.

Finally, these architectures manifest differently across biological domains:

- **Genomics/transcriptomics:** Highly standardized pipelines; dominated by single-agent and hybrid designs.
- **Proteomics/structural biology:** Heavy use of physics and molecular modeling tools; dominated by hybrid and multi-agent systems.
- **Single-cell/spatial biology:** Multimodal and heterogeneous; dominated by multi-agent and RAG-centric systems.
- **Drug discovery/therapeutic design:** Iterative design loops; rely on planner–critic–specialist ecosystems.
- **Biomedical reasoning/clinical agents:** Safety-critical; require ontology grounding, multi-modal fusion, and verifier modules.

Taken together, this taxonomy task type → agent architecture → agent roles → biological domain provides a compact structure for analyzing heterogeneous agentic systems and motivates the comparative evaluation presented in Section 5.

Critical analysis of applications across biological domains

Genomics and transcriptomics agents

Agentic LLM systems have gained rapid traction in genomics and transcriptomics, where workflows naturally involve multi-step reasoning, tool orchestration, and biological grounding. Early examples such as CRISPR-GPT [34] demonstrated that LLM controllers can design complete CRISPR experiments including sgRNA ranking, off-target assessment, delivery choices, and validation planning achieving expert-level performance in wet-lab benchmarks. BioDiscovery Agent [35] generalizes this idea to perturbation biology using planner–critic multi-agent loops, improving hit rates for both single-gene and combinatorial perturbations through literature retrieval and hypothesis refinement.

Beyond experimental design, several systems address regulatory inference and transcriptomic analysis. LLM4GRN [36] integrates LLM-derived regulatory priors with synthetic data generation to evaluate causal gene–TF relationships, outperforming conventional GRN inference baselines. SeqMate [37] shows that a single LLM agent can coordinate an end-to-end RNA-seq pipeline from FASTQ files through alignment, quantification, differential expression, and report generation substantially lowering the barrier for non-computational biologists. Downstream interpretation has also been improved through agentic designs. GeneAgent [38] incorporates a

self-verification loop, combining ontology lookup, evidence checking, and iterative reasoning to generate reliable gene-set annotations with markedly reduced hallucination rates. GenoTEX [39] provides a large-scale benchmark of over 900 gene-expression datasets and introduces "GenoAgents," a baseline multi-agent pipeline for dataset selection, preprocessing, statistical testing, and structured reporting.

More general-purpose frameworks are beginning to emerge. Biomni [40] combines planning, retrieval, code execution, and multi-omics tool-use within a single modular agent capable of gene prioritization, variant interpretation, and enrichment analysis. ChatNT [41] explores multimodal conversational reasoning over DNA, RNA, and protein sequences, while CompBioAgent [42] provides interactive LLM-driven exploration of scRNA-seq datasets, including clustering, marker identification, and annotation. Several systems focus on data integration and harmonization. BioNexusSentinel [43] illustrates how LLM-assisted coding accelerates the development of RNA-seq and network-visualization tools, whereas CellAtria [44] uses a planner–validator agent ecosystem to automate ingestion, QC, and metadata standardization at scale. For spatial transcriptomics, the Lightweight ST LLM Agent [45, 46] shows that even smaller models can support spatial-omics annotation via marker-based reasoning and atlas matching.

Taken together, genomics and transcriptomics represent one of the most mature domains for biological LLM agents. Across CRISPR design, perturbation modeling, GRN inference, end-to-end RNA-seq analysis, gene-set interpretation, and spatial-omics labeling, these systems demonstrate clear utility while underscoring ongoing challenges around grounding, error control, reproducibility, and generalization across cell types and species. See Supplementary Table S2 for a comparative overview of all systems in this domain.

Proteomics and structural biology agents

LLM-driven proteomics has progressed along two paths: (i) multimodal protein–language models that couple sequences, structures, and scientific text, and (ii) agentic systems that coordinate structure predictors, physics-based tools, and literature retrieval. Together, these approaches extend LLMs from descriptive tasks toward tool-integrated protein engineering and hypothesis generation. DrBioRight 2.0 exemplifies this trend by providing an LLM-guided assistant that automates routine sequence analysis and annotation through modular tool invocation [47].

Multimodal models such as ProteinGPT and ProtLLM integrate ESM-2 or ProtST encoders [48, 49], AlphaFold-derived representations [50], and protein-specific corpora to support Q&A, property prediction, and protein–text retrieval [33, 51]. These systems outperform general LLMs but remain non-agentic, lacking planning, external tool-use, and verification. X-LoRA introduces adapter-based experts for protein mechanics and molecular design [52], improving scientific reasoning while still operating as a single-shot model.

Agentic proteomics systems incorporate explicit roles and iterative control loops. ProteinHypothesis [53] employs BioAgent, StrucAgent, and Critic modules with retrieval-augmented generation and physics-aware checks to propose and rank candidate structural hypotheses. ProtAgents [54] implements a multi-agent AutoGen workflow in which GPT-4 coordinates OmegaFold, ProteinForceGPT, and ProDy to design proteins and analyze folds. TourSynbio-Search [55] focuses on scientific search, routing InternLM2-7B queries across UniProt, PDB, and preprint servers for interpretable protein engineering recommendations. ProteinEngine [56] generalizes these patterns with a project–expert–presenter triad orchestrating AlphaFold2, ESMFold, DiffDock, and ProtENN. ProtFunAgent [57] demonstrates how homology retrieval, GO-constrained decoding, and self-verification reduce hallucinations during functional annotation. At the experimental frontier, Virtual Lab [58] uses role-defined agents (PI, ML specialist, immunologist, computational biologist) coordinating ESM, AlphaFold-Multimer, and Rosetta to design antibody variants; two of 92 candidates showed improved ELISA binding.

In summary, proteomics agents fall into three groups: (i) multimodal backbones for protein–language understanding (ProteinGPT, ProtLLM, and X-LoRA), (ii) multi-agent controllers integrating diverse structural and physics tools (ProteinHypothesis, ProtAgents, TourSynbio-Search, and ProteinEngine), and (iii) early physics-integrated discovery workflows with experimental validation (Virtual Lab and ProtFunAgent). These systems signal a shift toward tool-integrated, autonomous LLM-driven protein science. Extended comparisons are provided in Supplementary Table S3.

Single-cell and spatial biology agents

Single-cell and spatial transcriptomics present high-dimensional, multimodal, and multi-step analysis challenges that align naturally with agentic LLMs. Recent systems combine LLM reasoning with established toolchains for preprocessing, clustering, annotation, and spatial interpretation.

uHAF [59] uses GPT-4 with a curated ontology (uHAF-T) to harmonize inconsistent scRNA-seq labels, improving downstream classification while relying on marker and ontology quality. CellAgent [60] provides a transformer baseline for clustering and marker discovery using sequence–structure encoders. scAgent (GPTCell) [61] automates the full scRNA-seq pipeline QC, clustering, annotation, batch correction, trajectory analysis via GPT-4 and tools including Scanpy [62], CellTypist [63], scVI-tools [64], Enrichr [65], Harmony [66], and scIB [67]; its adaptive planning yields expert-level annotations but incurs higher runtime.

In spatial transcriptomics, QuST-LLM [68] coordinates SpaGCN [69], BayesSpace [70], Tangram [71], Cell2location [72], and Squidpy [73] to perform domain identification, deconvolution, and enrichment. SpatialAgent [74] extends this framework with multimodal reasoning and memory, integrating scRNA-seq, MERFISH, and Visium data through Scanpy, LIANA+ [75], CellChat [76], and marker databases to support niche annotation and ligand–receptor analysis; fine-grained subtype resolution remains challenging. CASSIA [77] provides a multi-agent system (annotator, validator, scorer, and reporter) for reference-free annotation using GPT-4o, Claude 3.5, or LLaMA-3.2, achieving large gains (12%–41%) across GTEx, Tabula Sapiens, and Human Cell Landscape. Marker extraction quality remains a key dependency. SpaCCC [78], although not agentic, demonstrates how LLM embeddings (e.g. scGPT) combined with graph-based features improve ligand–receptor inference.

Overall, single-cell and spatial agents show a strong shift toward LLMs as controllers for multimodal workflows. Systems such as uHAF, scAgent, CASSIA, QuST-LLM, and SpatialAgent demonstrate scalable and interpretable pipelines powered by structured reasoning and tight integration with established bioinformatics tools. Supplementary Table S4 provides extended details.

Drug discovery and therapeutic design agents

LLM agents are increasingly used to automate early-stage therapeutic design, replacing linear pipelines with multi-step reasoning loops

that retrieve biochemical evidence, generate or refine molecules, run external predictors, and iteratively update decisions. Across systems, a common pattern emerges: LLMs act as planners and interpreters, while docking engines, ADMET models, and molecular toolkits provide domain-specific computation.

DrugAgent [79] demonstrates this architecture clearly. Through a multi-agent framework, it generates code, invokes ADMET and HTS models, and debugs outputs using RDKit, ChemBERTa, and ESM [48, 80, 81]. It achieves strong ROC–AUC on PAMPA, HIV, and DAVIS, though its task space and reliance on baseline models remain limiting. CLADD [82] extends this idea with retrieval-centric coordination. Using GPT-4o, Galactica, and MolecularGPT with PubChem and PrimeKG, it performs captioning, drug–target prediction, and toxicity modeling. Its advantage is dynamic external retrieval, while sensitivity to heterogeneous biochemical data is a key constraint.

Several systems adopt a broader, end-to-end perspective. PharmAgents [83] approximates a virtual pharmaceutical workflow from target identification to lead optimization integrating UniProt, PDB, CROSSDOCKED, PLIP, and docking engines within a GPT-4o/DeepSeek framework. It provides explainable intermediate outputs and strong docking-based scores, though it stops before clinical stages. ChatMol Copilot [84] supports multimodal molecular modeling (ESMFold, ProteinMPNN, RDKit, Vina, and 3D visualization), enabling flexible protein design and docking in a single interface, albeit with substantial computational cost and dependence on model knowledge.

Other approaches probe generative creativity. "Hallucinations Can Improve LLMs in Drug Discovery" [85] shows that controlled expansion of molecular descriptions can enhance toxicity and BBB prediction across HIV, BBBP, Clintox, SIDER, and Tox21 when using Llama-3, Ministral-8B, ChemLLM-7B, and GPT-4o. Systems such as SwiftDossier [86] instead focus on automating literature-heavy tasks, generating structured target dossiers using mid-size open-source models (Mistral-7B, Qwen-1.5-7B, Orca-13B, and Gemma-7B) with UniProt, DrugBank, and Open Targets; accuracy remains dependent on source data.

Taken together, drug-discovery agents converge on shared principles: LLMs act as controllers for docking and ADMET tools, rely on retrieval for biochemical grounding, use multi-step refinement instead of one-shot predictions, and generate interpretable intermediates for expert review. Remaining challenges include dataset heterogeneity, model bias, inconsistent evaluation, and limited wet-lab integration, but progress across DrugAgent, CLADD, PharmAgents, ChatMol Copilot, and related systems suggests a clear trajectory toward increasingly autonomous therapeutic design. See Supplementary Table S5 for detailed comparisons.

Biomedical reasoning and clinical agents

Biomedical LLM agents span clinical text reasoning, decision support, wearable–sensor analytics, and multimodal imaging. Unlike earlier RAG-style models, recent systems incorporate agentic behaviors such as autonomous graph construction, structured reasoning loops, and multimodal tool integration (Supplementary Table S6).

Clinical text–reasoning agents such as MedGraphRAG [87] and Health-LLM [88] combine large biomedical corpora, EHR data (e.g. MIMIC-IV), and ontologies like UMLS to organize medical evidence. MedGraphRAG builds a hierarchical knowledge graph and uses LLM controllers (GPT-4, Gemini, and Llama-2/3) for retrieval, ranking, and fact-checking, improving citation precision versus standard RAG. Health-LLM uses similar loops for personalized disease prediction, where the agent retrieves patient-specific information and iteratively optimizes its reasoning. Although their emphases differ explainable graph reasoning versus personalized prediction both highlight the value of structured external resources.

A second set of systems targets consumer and wearable health. PhysioLLM [89] and HealthAlpaca [90] integrate wearable signals (heart rate, steps, and sleep) with LLM reasoning to generate personalized insights. PhysioLLM autonomously computes trends from Fitbit data; HealthAlpaca extends this using fine-tuned medical LLMs (MedAlpacaand PMC-Llama) with RAG, chain-of-thought, and self-consistency to improve prediction of stress and fatigue. Systems such as DrHouse [91], LIFT-PD [92], and LLaSA [93] incorporate multimodal sensor streams and support diagnostic or activity-aware reasoning. These models show promise for health coaching but face limits in dataset duration, self-report bias, and external validation. A related line of work extends agentic reasoning to continuous glucose monitoring (CGM) and intervention design. GlyRAG [94] introduces a context-aware retrieval-augmented framework for long-horizon blood glucose forecasting. Unlike conventional time-series models that treat CGM signals as purely numerical sequences [95, 96], GlyRAG employs an LLM as a contextualization agent to generate structured clinical summaries directly from glucose traces. Complementing forecasting, SenseCF [97, 98] explores counterfactual modeling with fine-tuned LLMs for health intervention design and sensor data augmentation. The framework leverages LLMs (GPT-4, BioMistral-7B, LLaMA-3.1-8B) to generate clinically plausible counterfactual explanations that identify minimal, behaviorally actionable changes capable of altering model predictions.

A third cluster focuses on medical imaging. TriMedLM [99] integrates 3D vision transformers with LLMs for CT/MRI VQA, retrieval, localization, and segmentation. Zeus [100] uses an LLM (Vicuna-Rad) to autonomously generate segmentation instructions for SAM-based backbones, enabling zero-shot union segmentation across modalities. LLM-CXR [101] pairs language models with VQ-GAN or TorchXRayVision encoders for report generation and CXR-VQA, achieving competitive AUROC and BLEU with relatively small architectures. Collectively, these agents show that iterative prompting can effectively guide visual models through multi-step interpretation workflows.

Across biomedical domains, common themes emerge: nearly, all systems rely on external tools (graphs, ontologies, wearable analytics, and segmentation models), use proactive querying and iterative refinement, and adopt evaluation schemes that vary by modality (accuracy/expert review for text, user studies for wearables, and Dice/AUROC/BLEU for imaging). Despite clear progress, challenges remain in dataset bias, limited clinical validation, latency, and regulatory alignment. Standardized benchmarks and real-world prospective testing are needed for reliable deployment. See Supplementary Table S6 for details.

Biomedical imaging and vision–language agents

Medical imaging has become a major frontier for agentic LLMs, driven by the need to couple visual perception with clinical reasoning and workflow constraints. Unlike symbolic omics tasks, imaging requires models to interpret high-dimensional visual data while interacting with external tools. Recent systems increasingly incorporate multi-step planning, tool invocation, and human-in-the-loop refinement.

MID-LLM [102] exemplifies large-scale agentic integration by combining federated learning, blockchain auditability, and LLM-based

contextual reasoning for MRI analysis. Smart-contract coordination and verification modules support privacy-preserving diagnosis and detection of malicious updates. RaDialog [103] brings interactive capabilities to radiology, pairing domain-specific image encoders with instruction-tuned LLMs to support report generation, clarification, and image-grounded dialog.

Other systems focus on perception tasks. The Lightweight Multimodal VQA model [104] and MedVisionLlama [105] enhance VQA and segmentation using LLM layers as semantic priors, improving few-shot generalization across modalities. Broader multimodal frameworks such as PaLM-E [106] and Med-PaLM/MultiMedQA [107] demonstrate how LVLMs can integrate image understanding and clinical text reasoning. Surveys such as [108] similarly highlight the emergence of workflow-aware imaging agents.

Despite progress, imaging agents remain sensitive to acquisition variability, require substantial compute, and risk clinically harmful hallucinations. Regulatory pathways for autonomous radiology assistance are also immature. Nonetheless, the convergence of perception, multi-step reasoning, and workflow integration suggests strong potential for future diagnostic assistants. Supplementary Table S7 provides comparisons.

Bioinformatics question answering and agentic bio-analysis systems

Bioinformatics question answering traditionally depends on curated databases and static pipelines. LLM agents are transforming this area by enabling systems that retrieve information, execute analyses, and coordinate specialized sub-agents (Supplementary Table S8).

BioAgent [109] decomposes queries into genomics-, coding-, and analysis-focused agents using BioStar QA data, EDAM ontologies, and BioContainer APIs. Despite using small models (e.g. Phi-3-mini), it can generate runnable workflows, though it struggles with complex programming tasks and depends heavily on nf-core conventions. TxGemma [110] extends agentic reasoning to therapeutics with 2B–27B Gemma-2 variants, combining reasoning and acting workflows for drug–target prediction and explanation tasks. BixBench [111] provides one of the first comprehensive benchmarks across genomics, proteomics, metabolomics, and comparative genomics, evaluating autonomy, tool collaboration, and refusal behavior. It highlights persistent weaknesses in statistical reasoning and multimodal integration.

Other systems apply agentic reasoning to molecular modeling or biomedical knowledge graphs. X-LoRA-Gemma [112] integrates RDKit, PyMOL, and PySCF within a multi-agent framework for molecular proposal and evaluation. Talk2Biomodels and Talk2KnowledgeGraph [113] combine LLMs with LangGraph and GraphRAG to reason over biomarker networks and SBML models using PrimeKG. Causal inference is automated by MRAgent [114], which uses GWAS data and tools like TwoSampleMR and MRlap to identify exposure–outcome relationships with multiple LLM backbones. FlowAgent [115] demonstrates full workflow automation for RNA-seq, orchestrating FastQC, Kallisto, Hisat2, Salmon, and HPC resources to perform adaptive QC and error recovery.

Overall, BioQA and agentic bio-analysis systems are moving beyond retrieval toward autonomous execution, multimodal data integration, and structured biological reasoning. Challenges persist in reliability, explanation fidelity, and standardized evaluation, but rapid progress across BioAgent, TxGemma, BixBench, Talk2Biomodels, MRAgent, and FlowAgent indicates a clear trajectory toward more interactive and tool-using computational assistants.

Cross-sectional comparison and evaluation

Performance benchmarks and system capabilities

Across the literature, biological LLM agents converge into a few recurring capability patterns. Figure 2 highlights six qualitative axes tool-use, planning, accuracy, robustness, novelty, and biological grounding that consistently differentiate mature systems from exploratory prototypes. Direct, same-dataset comparison of biological LLM-agent systems within a single domain is currently infeasible. Most surveyed agents are evaluated on heterogeneous datasets, proprietary benchmarks, or task-specific pipelines, and often differ substantially in toolchains, objectives, and execution environments (see Table 3). As a result, raw performance numbers are not directly comparable across systems, even within the same biological domain.

Genomic perturbation and RNA-seq agents (e.g. BioDiscovery Agent [35] and SeqMate [37]) perform strongest in tool integration and grounding, reflecting their reliance on well-established genomics software and benchmarked datasets. Proteomics and structural biology agents also show strong biological grounding through physics-based tools (ESM, AlphaFold, and MD), but display weaker long-horizon planning and variable robustness due to instability in folding or docking pipelines.

Single-cell and spatial agents (e.g. scAgent/GPTCell, QuST-LLM, CASSIA, and SpatialAgent) present the most balanced profiles. They integrate heterogeneous modalities, choose modules dynamically for clustering, deconvolution, enrichment, or communication analysis, and often incorporate ontologies or retrieval, yielding high robustness but sensitivity to annotation quality and batch effects. Drug-discovery agents emphasize creative exploration through multi-agent or ensemble workflows; limitations stem largely from docking, ADMET, and heterogeneous assay data rather than LLM reasoning alone. Imaging and clinical agents show strong biological grounding but lag in multi-agent coordination and workflow depth due to less mature tool-API ecosystems. We note that recent efforts such as BixBench and GenoTEX represent early steps toward standardized evaluation of biological LLM agents, but comprehensive, domain-spanning benchmarks remain an open challenge.

Agent type vs. biological task matrix

Despite architectural diversity, biological LLM systems show a clear division of labor across tasks (Table 2). Single-agent controllers dominate well-structured workflows such as RNA-seq (SeqMate) and gene-set annotation (GeneAgent), where the primary challenge is reliable tool orchestration rather than deep scientific debate. These designs are transparent and lightweight but less suited for complex multi-stage reasoning. Multi-agent architectures are favored for hypothesis-driven or open-ended tasks. Systems such as SpatialAgent, CASSIA, PharmAgents, BioDiscoveryAgent, and DrugAgent combine planner, analyst, critic, and verifier roles to improve error detection and biological grounding. These ecosystems loosely mirror real scientific teams and perform best in domains that require integrating diverse evidence: spatial niches, perturbation screens, or lead optimization.

RAG-centric and tool-centric agents are most effective when validated algorithms or knowledge bases carry much of the domain-specific burden. LLM4GRN [36], CLADD, ProtFunAgent, and Talk2

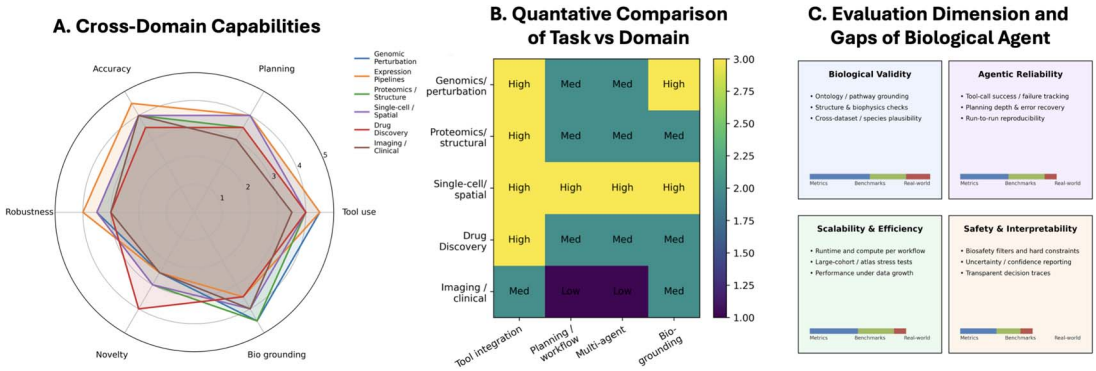

Figure 2 Cross-domain capabilities, task–domain alignment, and evaluation gaps of biological LLM agents. (A) Cross-domain capability radar chart. Comparison of seven representative biological domains shows substantial variability in accuracy, planning depth, tool-use reliability, robustness, novelty, and biological grounding. Spatial and proteomics agents tend to exhibit strong grounding, whereas drug discovery and imaging agents display higher variance in planning and robustness. (B) Heatmap of task requirements versus domain maturity. Domains differ in how strongly they depend on tool integration, planning or workflow orchestration, multi-agent collaboration, and biological grounding. Single-cell/spatial and drug-discovery agents consistently require the highest levels of planning and multi-tool coordination. (C) Evaluation dimensions and current gaps. Four major axes—biological validity, agentic reliability, scalability/efficiency, and safety/interpretability—highlight where present benchmarks fall short. While metrics exist for tool accuracy and basic grounding, real-world validation, reproducibility testing, and biosafety-aware evaluation remain sparse. Together, these panels illustrate the heterogeneous progress and persistent challenges across biological LLM-agent systems.

Table 2 Alignment between agent architectures and biological task domains. Multi-agent systems dominate exploratory, hypothesis-driven tasks (e.g. spatial biology and drug design), while single-agent and hybrid LLM–ML systems remain common in standardized pipelines such as RNA-seq and gene-set annotation.

Agent type	Genomic/ Perturbation	Expression pipelines	Proteomics/ Structure	Single-cell/ Spatial	Drug Discovery
Single-agent LLM	CRISPR-GPT, GeneAgent	SeqMate, lightweight RNA-seq agents	ProteinGPT, ProtLLM	SpaCCC (non-agentic), ChatNT	Few examples; mostly used as assistants
Multi-agent pipelines	BioDiscoveryAgent (planner–critic)	GenoAgents (planner, executor, validator)	Virtual Lab, ProtAgents	CASSIA, SpatialAgent, QuST-LLM	DrugAgent, PharmAgents (multi-role systems)
RAG-centric/tool-centric agents	LLM4GRN (GRNBoost2 + GAN)	Biomni, Talk2Biomodels	X-LoRA–Gemma, CLADD	SpatialAgent (knowledge tools), QuST-LLM	CLADD, Talk2KnowledgeGraph
Hybrid LLM + domain ML	GRNBoost2 + LLM reasoning	RNA-seq statistical tools (DESeq2, Kallisto)	AlphaFold, ESM2/IF1, DiffDock	Cell2location, Tangram, LIANA+	Cheminformatics pipelines (ChemBERTa, RDKit)

Table 3 Within-domain comparison of representative biological LLM-agent systems using capability- and evaluation-based criteria, rather than shared datasets.

Domain	Representative agents	Agentic capabilities	Evaluation type	Reported benchmarks
Genomics	GeneAgent, GenoTEX, SeqMate	Tool-use, planning, verification	Quantitative + expert review	Task-specific datasets
Proteomics	ProtAgents, Protein Hypothesis	Multi-agent, physics-aware	Simulation + expert scoring	Structure/mechanics tasks
Single-cell/Spatial	scAgent, SpatialAgent, CASSIA	Planning, memory, tool orchestration	ARI/NMI, expert annotation	Atlas-specific datasets
Drug Discovery	DrugAgent, PharmAgents, CLADD	Multi-agent collaboration	ROC–AUC, validity rates	ADMET/DTI benchmarks
Clinical/QA	MedGraphRAG, DrHouse	RAG, iterative refinement	Accuracy, citation precision	Clinical QA datasets

Biomodels rely heavily on GRN inference tools, cheminformatics stacks, BLAST/PDB, or large biomedical KGs. Here, the LLM primarily coordinates and interprets rather than replaces domain methods. Finally, hybrid LLM–ML systems dominate physics-intensive or statistical tasks, including protein structure prediction (AlphaFold and ESMFold), spatial deconvolution (Cell2location and Tangram),

Table 4 Operational definitions of evaluation dimensions commonly used in biological LLM-agent systems, and their typical quantitative or semi-quantitative proxies reported in prior work.

Evaluation dimension	Operational definition	Typical proxies used in practice
Biological validity	Consistency of agent outputs with established biological knowledge, mechanisms, or experimental evidence.	Pathway or GO-term overlap, agreement with curated databases, expert review, experimental validation.
Agentic reliability	Stability and coherence of multi-step reasoning and decisions across runs, prompts, or tool invocations.	Self-consistency checks, disagreement rates, plan stability, success rate of tool execution.
Grounding and faithfulness	Degree to which generated claims are supported by explicit evidence from tools, databases, or retrieval sources.	Citation verification, tool-call tracing, provenance-aware reasoning, evidence attribution accuracy.
Robustness	Sensitivity of agent behavior to perturbations in inputs, prompts, data noise, or intermediate states.	Prompt ablations, noise injection, dataset subsampling, stress testing under missing or conflicting evidence.
Reproducibility	Ability to reproduce identical or equivalent outputs under controlled execution settings.	Deterministic tool execution, logged workflows, environment versioning, replayable pipelines.

and docking (DiffDock). In these settings, LLMs contribute planning, explanation, and workflow assembly, while domain-specific models provide accuracy and stability. No single architecture is universally optimal; domains gravitate toward designs that match their task structure and tool landscape.

Operational definitions of evaluation dimensions

While biological LLM-agent systems are evaluated across a wide range of tasks and domains, there is currently no standardized benchmark suite that supports direct, cross-system comparison. Instead, prior work relies on a set of recurring evaluation dimensions that are operationalized using task-specific quantitative or semi-quantitative proxies. To clarify how these dimensions are measured in practice, we summarize their operational definitions and commonly used proxies in Table 4.

Importantly, our analysis does not assign unified scores to systems, but rather synthesizes how evaluation is performed across the literature, enabling structured comparison of evaluation practices rather than raw performance.

Evaluation dimensions and gaps

Despite rapid progress, biological LLM agents still lack a unified evaluation framework. Most studies measure only *end-task accuracy* (e.g. AUROC for drug–target prediction, ARI/NMI for scRNA-seq, ROUGE/BERTScore for protein or clinical text), while largely overlooking whether the *agentic process* itself is reliable, reproducible, or biologically sound. A key gap is insufficient assessment of **biological validity**. Agents for GRN inference, spatial annotation, CRISPR design, or protein function prediction rarely evaluate ontology consistency, pathway coherence, structural plausibility, or cross-modal agreement. Only a few systems (e.g. Virtual Lab and CRISPR-GPT) report any wet-lab validation. A second gap is limited reporting of **agentic competence**. Planning behavior, tool-use correctness, execution failures, and reproducibility are seldom quantified, even though many pipelines call fragile tools, such as BLAST, HMMER, AlphaFold, Scanpy, or docking engines where silent errors are common. Third one concerns **scalability and robustness**. Few agents are stress-tested on atlas-scale single-cell data, proteome-wide structure sets, or long genomic loci. Multimodal tasks still lack shared benchmarks, and safety, interpretability, and uncertainty estimation remain weakly developed. Domain-specific failure modes and mitigation strategies are summarized in Supplementary Table S2.

Across surveyed studies, most evaluations rely on proxy-based or task-specific metrics, underscoring the need for standardized, agent-aware benchmarks tailored to biological reasoning and multi-step workflows. Described challenges point toward four needed evaluation pillars: (i) *Biological validity* (ontology/pathway grounding, structural plausibility); (ii) *Agentic reliability* (planning, tool success rate, error recovery, and hallucination control); (iii) *Scalability and efficiency* (runtime, cost, and large-dataset performance); and (iv) *Safety and interpretability* (biosafety alignment, transparent reasoning, and uncertainty estimates). Adopting such criteria will enable more comparable and trustworthy assessment of biological LLM agents.

Cross-domain failure modes and design gaps

Beyond differences in application domains, biological LLM-agent systems share recurring failure patterns that arise from data heterogeneity, fragile tool interfaces, and underspecified biological priors. We synthesize these issues across domains in Supplementary Table S2 and illustrate typical error propagation in multi-step pipelines in Supplementary Fig. S1. Across surveyed systems, failures often compound over stages noise or missing metadata can trigger retrieval misalignment, which, in turn, leads to incorrect intermediate hypotheses, tool misuse, and amplified errors in multi-agent settings.

Domain-specific failure modes exhibit consistent signatures. In *genomics and transcriptomics*, agents frequently mis-handle gene or transcript identifiers (e.g. ambiguous symbols and assembly/version mismatches), producing incorrect differential signals and downstream pathway interpretations. In *proteomics and protein design*, unconstrained generation may yield non-physical sequences or structures without explicit structural constraints, resulting in failed folding, docking, or unrealistic candidates. In *single-cell and spatial biology*, coordinate/metadata inconsistencies and platform-dependent resolution can cause misalignment between spatial context and labels, leading to erroneous niche or domain assignments. In *drug discovery*, agents may propose chemically invalid or non-synthesizable molecules when generation is weakly constrained, wasting expensive downstream evaluation. In *clinical and biomedical QA*, incomplete corpora or shallow retrieval can produce overconfident but unsupported responses, raising safety and trust concerns. Finally, in *workflow automation*, silent tool invocation errors, environment drift, and brittle

file formats can yield non-reproducible analyses without explicit provenance and validation.

These observations motivate concrete design gaps that persist across domains: (i) ontology- and version-aware grounding (e.g. robust ID mapping and reference tracking), (ii) verification layers that validate intermediate outputs (e.g. constraint checks, provenance-aware citations, and consistency tests), and (iii) uncertainty-aware decision-making to avoid brittle single-shot commitments under ambiguous evidence.

Technical challenges and open problems

Across domains, biological LLM agents remain limited by three cross-cutting issues: (i) data alignment and hallucination control, (ii) interpretability and validation, and (iii) scalability and integration with laboratory or clinical systems (Supplementary Table S2). These challenges stem from both LLM limitations and the complexity of biological data, tools, and workflows.

Data alignment and hallucination

Biological hallucination, plausible but incorrect genes, sequences, pathways, or explanations remains one of the most serious risks. Retrieval-heavy systems such as CLADD, MedGraphRAG, and MRAgent often generate unsupported statements when knowledge graphs are incomplete. Heterogeneous datasets (GWAS, EHRs, single-cell, and spatial) and inconsistent metadata further worsen misalignment. Multi-step agents such as FlowAgent, LLaSA, or TriMedLM can propagate subtle preprocessing or tool-chain errors across entire workflows. While ontology-aware normalization and standardized schemas are commonly proposed to address data alignment, their effectiveness depends on the completeness and consistency of upstream metadata, which is often missing or noisy in practice. Over-reliance on automated alignment may introduce silent errors that propagate through agentic workflows, particularly when versioning or reference mismatches are not explicitly tracked (see Supplementary Table S2). Design-oriented systems (DrugAgent, X-LoRA-Gemma, and ProteinGPT derivatives) also risk producing non-physical proteins or synthetically implausible molecules. Standardized schemas, stronger retrieval filters, and validation layers are needed to prevent such cascades (Supplementary Fig. S1). Tool augmentation and retrieval-based grounding can reduce hallucinations, but may also amplify errors when retrieved evidence is incomplete, outdated, or weakly relevant. In such cases, agents may produce confidently incorrect reasoning chains that are difficult to detect without explicit verification or uncertainty estimation mechanisms.

Interpretability and validation

Most agents provide chain-of-thought or citation-style explanations, but these often reflect linguistic fluency rather than mechanistic grounding. High-stakes tasks drug toxicity modeling (CLADD), radiology reasoning (LLM-CXR and RaDialog), and causal inference (MRAgent) require explicit evidence linkage, uncertainty estimates, and biological justification. Tool invocation adds opacity: agents frequently run Kallisto, AutoDock Vina, TwoSampleMR, or docking tools without verifying outputs or detecting silent failures, and even systems like FlowAgent only partially address this. Benchmarking remains sparse; while BixBench offers early multi-tool evaluation, many areas (spatial transcriptomics and protein engineering) still lack standardized, experimentally grounded tests. Although chain-of-thought prompting and intermediate state logging improve transparency, these signals do not guarantee biological correctness and may give a false sense of interpretability. Moreover, exposing detailed reasoning traces can increase cognitive load for users and complicate validation in high-stakes biomedical settings. As a result, expert oversight remains essential, limiting autonomy, and reproducibility.

Scalability and integration with experiments

Most systems remain proof-of-concept. Large multimodal models (TriMedLM and PaLM-E) require substantial compute, while multi-agent designs multiply cost and fragility. Biological pipelines depend on numerous external tools, databases, and environment-specific configurations; minor version drift frequently breaks workflows, as noted in FlowAgent, BioAgents, and Talk2Biomodels. Integration with wet-lab or clinical settings is minimal: few agents interact with robotic labs or decision-support systems, hindering true closed-loop pipelines (design → execute → learn). But scaling agentic systems across large datasets or institutions often requires complex toolchains and infrastructure, increasing engineering overhead and failure points. Multi-agent parallelization may improve throughput but also raises coordination challenges and the risk of error amplification across interacting agents (Supplementary Fig. S1). Privacy, biosecurity, and regulatory constraints further limit deployment, and most surveyed systems lack explicit safety or auditability mechanisms. Addressing these limitations is critical for reliable and safe biological LLM agents.

Taken together, these challenges highlight that proposed solutions for biological LLM agents involve non-trivial trade-offs between automation, reliability, interpretability, and scalability. Without careful design, mitigation strategies may shift rather than eliminate failure modes, reinforcing the need for verification layers, uncertainty-aware reasoning, and domain-informed evaluation protocols.

Opportunities and future trajectories

Toward adaptive and interactive bio-agents

Most current agents operate in a single-shot or short-horizon mode. A key opportunity is building adaptive systems that maintain state about the user, dataset, and workflow, enabling them to track past decisions, surface uncertainty, or request missing metadata rather than hallucinate around gaps. Exposing internal plans as editable structures would let experts adjust modules (e.g. changing DE methods or docking engines) with updates propagated through the pipeline. Lightweight preference learning remembering tools, ontologies, QC rules, or reporting styles could further support natural human–AI collaboration without retraining. Realizing adaptive and interactive bio-agents will require persistent memory modules, uncertainty-aware decision policies, and feedback signals derived from downstream analysis outcomes or user correction, as demonstrated in recent agentic systems for single-cell and spatial analysis [61, 74]. In practice, such feedback is often sparse, delayed, or noisy, limiting stable adaptation across datasets and tasks. Moreover, uncontrolled adaptation risks reinforcing spurious hypotheses or dataset-specific biases, as highlighted by recent benchmark analyses of agent failure behaviors [111], motivating constrained updates, and explicit verification layers.

Autonomous discovery and closed-loop learning

Early examples already hint at closed-loop design–evaluate–refine cycles in protein engineering, CRISPR design, and therapeutics. The next step is connecting these loops to experimental infrastructure under strict safety controls so that synthesis, screening, and imaging results feed back into agent reasoning. Such systems could maintain evolving hypotheses, prioritize experiments that reduce uncertainty, and adapt to real-time data. Analogous opportunities exist in clinical contexts adaptive trial design, surveillance, and risk forecasting provided drift, bias, and safety are continually monitored. Closed-loop biological agents rely on iterative feedback from experiments, simulations, or downstream analyses to refine hypotheses and decisions, as demonstrated in active learning frameworks for transcriptomics-driven phenotype modulation and emerging general-purpose biomedical agents [40, 74, 116]. Recent surveys of agentic science further highlight the promise of autonomous discovery through closed-loop reasoning and experimentation [31]. However, biological feedback is often expensive, delayed, or noisy, and naively closing the loop can amplify early reasoning errors or overfit surrogate objectives. Careful design of feedback signals, safety constraints, and human-in-the-loop checkpoints is therefore essential for stable and biologically meaningful learning.

Multi-agent collaboration and workflow automation

Multi-agent designs are emerging across domains, yet most are bespoke. A major opportunity is standardized workflow substrates in which agents, tools, and datasets form modular, composable components (e.g. "QC → normalization → clustering → annotation → enrichment"). These would support automated tool selection, provenance tracking, and fully inspectable analyses. Richer collaboration patterns modality-specialized, task-specialized, or domain-specific agents could negotiate proposals, challenge assumptions, and escalate inconsistencies to human experts. Robust monitoring and replayable execution traces will be essential for scientific and regulatory auditing. Together, these developments support trustworthy multi-agent ecosystems capable of automating complex discovery pipelines. Multi-agent collaboration enables modular specialization and parallel reasoning, as demonstrated in drug discovery and clinical decision-support systems, including recent clinical triage agents [79, 83, 117]. TriageAgent, in particular, shows how role-specialized agents can improve decision quality through structured coordination. However, multi-agent systems introduce coordination and arbitration challenges, and without explicit consensus or verification mechanisms, interacting agents may diverge or amplify errors, as observed in recent benchmark studies [111].

Overall, while these future directions are technically plausible, their realization requires careful balancing of automation, reliability, and biological grounding. Naively scaling agent autonomy without feasibility-aware constraints risks shifting, rather than resolving, existing failure modes, reinforcing the importance of verification, uncertainty estimation, and domain-informed evaluation.

Conclusion

LLM-based agents increasingly act as a reasoning and orchestration layer above traditional bioinformatics tools. Across genomics, proteomics, spatial biology, drug design, and clinical informatics, recent systems show that even early-stage agents can coordinate multi-step workflows, integrate heterogeneous evidence, and assist exploratory scientific analysis. This marks a shift from passive text models to systems that structure analyses, choose tools, and diagnose failures.

Yet capabilities remain uneven. Biological grounding is inconsistent, tool-use can be brittle, and reproducibility is often difficult to guarantee. Agentic systems are promising but not mature enough for autonomous deployment in laboratory or clinical settings. The literature suggests a clear trajectory: as models incorporate stronger domain knowledge, reliable tool interfaces, and transparent execution environments, agentic LLMs are likely to evolve from experimental prototypes into stable components of computational biology. Their long-term impact will depend on rigorous validation, safety-aware design, and alignment with real scientific practice foundations necessary for transforming LLM agents into reliable collaborators in biological discovery.

Key Points

- Large language models are rapidly evolving into agentic systems that plan, execute tools, and reason across multimodal biological data.
- We systematically survey >60 LLM-based agents spanning genomics, proteomics, single-cell and spatial biology, drug discovery, bioinformatics workflows, and biomedical applications.
- We propose a unified taxonomy of biological LLM agents based on task type, architecture, and agentic capabilities, such as planning, tool-use, memory, and grounding.
- We identify shared technical challenges including data alignment, hallucination control, interpretability, and reproducibility and summarize emerging mitigation strategies.
- We outline future trajectories toward adaptive bio-agents, closed-loop discovery systems, and multi-agent collaboration frameworks for trustworthy biological computation.

Acknowledgements

We gratefully acknowledge the funders' support and resources that made this research possible.

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Conflicts of interest

None declared.

Funding

This work was supported in part by Virginia Tech, the Department of Computer Science, and the U.S. National Science Foundation (NSF) under Awards nos 2125798, 2344169, and 2319522.

Data availability

This article is a methodological and literature-based review and does not generate new datasets. All data used for analysis consists of publicly available manuscripts, software repositories, and benchmark resources, which are cited throughout the text and listed in Supplementary Tables S2–S8.

References

1. Guo H, Xing X, Zhou Y. *et al.* A survey of large language model for drug Research and Development. *IEEE Access* 2025;**13**:51110–29. https://doi.org/10.1109/ACCESS.2025.3552256
2. Luo Y, Zhang J, Fan S. *et al.* BioMedGPT: an open multimodal large language model for biomedicine. *IEEE J Biomed Health Inform* 2024;**30**:981–92.
3. Abir AR, Tahmid MT, Rayan RI. *et al.* DeepRNA-twist: language-model-guided RNA torsion angle prediction with attention-inception network. *Brief Bioinform* 2025;**26**. https://doi.org/10.1093/bib/bbaf199
4. Wang C, Li M, He J. *et al.* A survey for large language models in biomedicine. *Artif Intell Med* 2025;**170**:103268. https://doi.org/10.1016/j.artmed.2025.103268
5. Bi Z, Dip SA, Hajialigol D. *et al.* AI for biomedicine in the era of large language models. arXiv preprint arXiv:240315673. 2024.
6. Wang B, Xie Q, Pei J. *et al.* Pre-trained language models in biomedical domain: a systematic survey. *ACM Comput Surv* 2023;**56**:1–52.
7. Abir AR, Toki Tahmid M, Bayzid MS. BioLLMNet: enhancing RNA-interaction prediction with a specialized cross-LLM transformation network. *Brief Bioinform* 2025;**26**. https://doi.org/10.1093/bib/bbaf549
8. Lan W, Tang Z, Liu M. *et al.* The large language models on biomedical data analysis: a survey. *IEEE J Biomed Health Inform* 2025;**29**:4486–97. https://doi.org/10.1109/JBHI.2025.3530794
9. Zhang Q, Ding K, Lv T. *et al.* Scientific large language models: a survey on biological & chemical domains. *ACM Comput Surv* 2025;**57**:1–38.
10. Moor M, Banerjee O, Abad ZSH. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* 2023;**616**:259–65.
11. Sultan KA, Hisham MHH, Orkild B. *et al.* HAMIL-QA: hierarchical approach to multiple instance learning for atrial LGE MRI quality assessment. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Morocco: Springer; 2024. p. 275–84. https://doi.org/10.1007/978-3-031-72378-0_26
12. Che Y, Rafsani F, Shah J. *et al.* AnoFPDM: anomaly detection with forward process of diffusion models for brain MRI. In: *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 1113–22. Arizona, 2025.
13. Rafsani F, Sheth D, Che Y. *et al.* Leveraging multi-modal foundation model image encoders to enhance brain MRI-based headache classification. *Sci Rep* 2025;**15**:33256. https://doi.org/10.1038/s41598-025-18507-8
14. Rafsani F, Shah J, Chong CD. *et al.* DinoAtten3D: slice-level attention aggregation of DinoV2 for 3D brain MRI anomaly classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3544–53. Hawaii, USA, 2025.
15. Rafsani F, Sheth D, Che Y. *et al.* Using large-scale contrastive language-image pre-training to maximize brain MRI-based headache classification (P4–12.007). In: *Neurology*, Vol. **104**, p. 2728. MD: Lippincott Williams & Wilkins Hagerstown, 2025.
16. Yi Z, Xiao T, Albert MV. A survey on multimodal large language models in radiology for report generation and visual question answering. *Information*. 2025;**16**:136.
17. Hisham MHH, Elhabian S, Adluru G. *et al.* Family of deep image prior networks for accelerated 3D LGE-MRI acquisition with enhanced reconstruction. In: *Medical Imaging with Deep Learning*. Salt Lake City, UT, USA, 2025.
18. Ali S, Qadri YA, Ahmad K. *et al.* Large language models in genomics—a perspective on personalized medicine. *Bioengineering* 2025;**12**:440.
19. Dip SA, Ma D, Zhang L. DeepAge: harnessing deep neural network for epigenetic age estimation from DNA methylation data of human blood samples. In: *Proceedings of the AAAI Symposium Series*, Vol. **4**, pp. 267–74. Arlington, Virginia, 2024.
20. Abir AR, Dip SA, Zhang L. UnCOT-AD: unpaired cross-omics translation enables multi-omics integration for Alzheimer's disease prediction. *Brief Bioinform* 2025;**26**. https://doi.org/10.1093/bib/bbaf438
21. Dip SA, Zafor A, Paul BK. *et al.* LLM4Cell: a survey of large language and agentic models for single-cell biology. arXiv preprint arXiv:251007793. 2025.
22. Abir AR, Dip SA, Zhang L. CFM-GP: unified conditional flow matching to learn gene perturbation across cell types. arXiv preprint arXiv:250808312. 2025.
23. Ahmed S, Emon MI, Moumi NA. *et al.* ProtAlign-ARG: antibiotic resistance gene characterization integrating protein language models and alignment-based scoring. *Sci Rep* 2025;**15**:30174.
24. Wang L, Ma C, Feng X. *et al.* A survey on large language model based autonomous agents. *Front Comp Sci* 2024;**18**:186345.
25. Cao Y, Zhao H, Cheng Y. *et al.* Survey on large language model-enhanced reinforcement learning: concept, taxonomy, and methods. *IEEE Trans Neural Networks Learn Syst* 2024;**36**:9737–57. https://doi.org/10.1109/TNNLS.2024.3497992
26. Wu Q, Bansal G, Zhang J. *et al.* Autogen: enabling next-gen LLM applications via multi-agent conversations. In: *First Conference on Language Modeling*. Philadelphia, Pennsylvania, United States, 2024.
27. Zhang Z, Dai Q, Bo X. *et al.* A survey on the memory mechanism of large language model-based agents. *ACM Trans Inform Syst* 2025;**43**:1–47.
28. Ruan W, Lyu Y, Zhang J. *et al.* Large language models for bioinformatics. *Quant Biol* 2026;**14**:e70014. https://doi.org/10.1002/qub2.70014
29. Yang T, Xiao Y, Bao Z. *et al.* The rise and potential opportunities of large language model agents in bioinformatics and biomedicine. *Brief Bioinform* 2025;**26**. https://doi.org/10.1093/bib/bbaf601
30. Chen K, Wang P, Yu Y. *et al.* Large language model-based data science agent: a survey. arXiv preprint arXiv:250802744. 2025.
31. Wei J, Yang Y, Zhang X. *et al.* From AI for science to agentic science: a survey on autonomous scientific discovery. arXiv preprint arXiv:250814111. 2025.
32. Zhou J, Li H, Chen S. *et al.* Large language models in biomedicine and healthcare. *Npj Artif Intell* 2025;**1**:44.
33. Xiao Y, Sun E, Jin Y. *et al.* ProteinGPT: multimodal llm for protein property prediction and structure understanding. arXiv preprint arXiv:240811363. 2024.
34. Qu Y, Huang K, Yin M. *et al.* CRISPR-GPT for agentic automation of gene-editing experiments. *Nat Biomed Eng* 2026;**10**:245–58.
35. Roohani Y, Lee A, Huang Q. *et al.* BioDiscoveryAgent: an AI agent for designing genetic perturbation experiments. arXiv preprint arXiv:240517631. 2024.

36. Afonja T, Sheth I, Binkyte R. *et al.* LLM4GRN: discovering causal gene regulatory networks with LLMs–evaluation through synthetic data generation. arXiv preprint arXiv:241015828. 2024.
37. Mondal D, Inamdar A. SeqMate: a novel large language model pipeline for automating RNA sequencing. arXiv preprint arXiv:240703381. 2024.
38. Wang Z, Jin Q, Wei C. *et al.* GeneAgent: self-verification language agent for gene set knowledge discovery using domain databases. arXiv. arXiv preprint arXiv:240516205. 2024.
39. Liu H, Chen S, Zhang Y. *et al.* GenoTEX: an LLM agent benchmark for automated gene expression data analysis. arXiv preprint arXiv:240615341. 2024.
40. Huang K, Zhang S, Wang H. *et al.* Biomni: a general-purpose biomedical AI agent. bioRxiv. 2025.
41. Richard G, de Almeida BP, Dalla-Torre H. *et al.* ChatNT: a multimodal conversational agent for DNA, RNA and protein tasks. bioRxiv. 2024;2024–04.
42. Zhang H, Sun YH, Hu W. *et al.* CompBioAgent: an LLM-powered agent for single-cell RNA-seq data exploration. bioRxiv. 2025;2025–03.
43. Matzko RO, Konur S. BioNexusSentinel: a visual tool for bioregulatory network and cytohistological RNA-seq genetic expression profiling within the context of multicellular simulation research using ChatGPT-augmented software engineering. *Bioinform Adv* 2024;**4**. https://doi.org/10.1093/bioadv/vbae046
44. Nouri N, Artzi R, Savova V. An agentic AI framework for ingestion and standardization of single-cell RNA-seq data analysis. bioRxiv. 2025;2025–07.
45. Dip SA, Zhang L. Can lightweight LLM agents improve spatial transcriptomics annotation? bioRxiv. 2025;2025–11.
46. Dip SA, Zhang L. NicheAgent: LLM-guided zero-shot niche identification for spatial transcriptomics. bioRxiv. 2025;2025–12.
47. Liu W, Li J, Tang Y. *et al.* DrBioRight 2.0: an LLM-powered bioinformatics chatbot for large-scale cancer functional proteomics analysis. *Nat Commun* 2025;**16**:2256.
48. Lin Z, Akin H, Rao R. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30.
49. Xu M, Yuan X, Miret S. *et al.* ProtST: multi-modality learning of protein sequences and biomedical texts. In: *International Conference on Machine Learning*, pp. 38749–67. Hawaii, USA: PMLR, 2023.
50. Jumper J, Evans R, Pritzel A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
51. Zhuo L, Chi Z, Xu M. *et al.* ProtLLM: an interleaved protein-language LLM with protein-as-word pre-training. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8950–63. Bangkok, Thailand, 2024.
52. Buehler EL, Buehler MJ. X-LoRA: mixture of low-rank adapter experts, a flexible framework for large language models with applications in protein mechanics and molecular design. *APL Mach Learn* 2024;**2**. https://doi.org/10.1063/5.0203126
53. Bazgir A, Madugula RCP, Zhang Y. Proteinhypothesis: a physics-aware chain of multi-agent rag LLM for hypothesis generation in protein science. In: *Towards Agentic AI for Science: Hypothesis Generation, Comprehension, Quantification, and Validation*, 2025.
54. Ghafarollahi A, Buehler MJ. ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery* 2024;**3**:1389–409. https://doi.org/10.1039/d4dd00013g
55. Liu Y, Chen Z, Wang YG. *et al.* Toursynbio-search: a large language model driven agent framework for unified search method for protein engineering. In: *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 5395–400. IEEE, 2024.
56. Shen Y, Lv O, Zhu H. *et al.* Proteinengine: empower LLM with domain knowledge for protein engineering. In: *International Conference on Artificial Intelligence in Medicine*, pp. 373–83. Springer, 2024. https://doi.org/10.1007/978-3-031-66538-7_37.
57. Dip SA, Choy JS, Zhang L. ProtFunAgent: agentic LLM cascades for low-resource protein function gap-filling via homology RAG and Ontology-constrained decoding. In: *NeurIPS 2025 Workshop on Efficient Reasoning*, 2025.
58. Swanson K, Wu W, Bulaong NL. *et al.* The virtual lab: AI agents design new SARS-CoV-2 nanobodies with experimental validation. bioRxiv. 2024;2024–11.
59. Bian H, Chen Y, Wei L. *et al.* uHAF: a unified hierarchical annotation framework for cell type standardization and harmonization. *Bioinformatics* 2025;**41**. https://doi.org/10.1093/bioinformatics/btaf149
60. Xiao Y, Liu J, Zheng Y. *et al.* Cellagent: an LLM-driven multi-agent framework for automated single-cell data analysis. arXiv preprint arXiv:240709811. 2024.
61. Mao Y, Mi Y, Liu P. *et al.* Scagent: universal single-cell annotation via a LLM agent. arXiv preprint arXiv:250404698. 2025.
62. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:15.
63. Domínguez, Conde C, Xu C, Jarvis LB. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* 2022;**376**. https://doi.org/10.1126/science.abl5197
64. Gayoso A, Lopez R, Xing G. *et al.* A python library for probabilistic analysis of single-cell omics data. *Nat Biotechnol* 2022;**40**:163–6. https://doi.org/10.1038/s41587-021-01206-w
65. Kuleshov MV, Jones MR, Rouillard AD. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;**44**:W90–7. https://doi.org/10.1093/nar/gkw377
66. Korsunsky I, Millard N, Fan J. *et al.* Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**:1289–96. https://doi.org/10.1038/s41592-019-0619-0
67. Luecken MD, Büttner M, Chaichoompu K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 2022;**19**:41–50. https://doi.org/10.1038/s41592-021-01336-8
68. Huang CH. QuST-LLM: integrating large language models for comprehensive spatial transcriptomics analysis. arXiv preprint arXiv:240614307. 2024.
69. Hu J, Li X, Coleman K. *et al.* SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat Methods* 2021;**18**:1342–51. https://doi.org/10.1038/s41592-021-01255-8
70. Zhao E, Stone MR, Ren X. *et al.* BayesSpace enables the robust characterization of spatial gene expression architecture in tissue sections at increased resolution. bioRxiv. 2020;2020–09.
71. Biancalani T, Scalia G, Buffoni L. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nat Methods* 2021;**18**:1352–62. https://doi.org/10.1038/s41592-021-01264-7
72. Kleshchevnikov V, Shmatko A, Dann E. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat Biotechnol* 2022;**40**:661–71.
73. Palla G, Spitzer H, Klein M. *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nat Methods* 2022;**19**:171–8.
74. Wang H, He Y, Coelho PP. *et al.* SpatialAgent: an autonomous AI agent for spatial biology. bioRxiv. 2025;2025–04.
75. Dimitrov D, Schäfer PSL, Farr E. *et al.* LIANA+ provides an all-in-one framework for cell–cell communication inference.

Nat Cell Biol 2024;**26**:1613–22. https://doi.org/10.1038/s41556-024-01469-w
76. Jin S, Plikus MV, Nie Q. CellChat for systematic analysis of cell–cell communication from single-cell transcriptomics. *Nat Protoc* 2025;**20**:180–219. https://doi.org/10.1038/s41596-024-01045-4
77. Xie E, Cheng L, Shireman J. *et al.* CASSIA: a multi-agent large language model for reference free, interpretable, and automated cell annotation of single-cell RNA-sequencing data. bioRxiv. 2024;2024–12.
78. Ji B, Xu L, Peng S. SpaCCC: large language model-based cell-cell communication inference for spatially resolved transcriptomic data. bioRxiv. 2024;2024–02.
79. Liu S, Lu Y, Chen S. *et al.* Drugagent: automating AI-aided drug discovery programming through LLM multi-agent collaboration. arXiv preprint arXiv:241115692. 2024.
80. Bento AP, Hersey A, Félix E. *et al.* An open source chemical structure curation pipeline using RDKit. *J Chem* 2020;**12**:51.
81. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:201009885. 2020.
82. Lee N, De Brouwer E, Hajiramezanali E. *et al.* RAG-enhanced collaborative LLM agents for drug discovery. arXiv preprint arXiv:250217506. 2025.
83. Gao B, Huang Y, Liu Y. *et al.* Pharmagents: building a virtual pharma with large language model agents. arXiv preprint arXiv:250322164. 2025.
84. Sun J, Li A, Deng Y. *et al.* ChatMol Copilot: an agent for molecular Modeling and computation powered by LLMs. In: *Proceedings of the 1st Workshop on Language+ Molecules (L+ M 2024)*, pp. 55–65, 2024.
85. Yuan S, Färber M. Hallucinations can improve large language models in drug discovery. arXiv preprint arXiv:250113824. 2025.
86. Fossi G, Boulaimen Y, Outemzabet L. *et al.* SwiftDossier: tailored automatic dossier for drug discovery with LLMs and agents. arXiv preprint arXiv:240915817. 2024.
87. Wu J, Zhu J, Qi Y. *et al.* Medical graph rag: towards safe medical large language model via graph retrieval-augmented generation. arXiv preprint arXiv:240804187. 2024.
88. Jin M, Yu Q, Shu D. *et al.* Health-LLM: personalized retrieval-augmented disease prediction system. arXiv preprint arXiv:240200746. 2024.
89. Fang CM, Danry V, Whitmore N. *et al.* PhysioLLM: supporting personalized health insights with wearables and large language models. In: *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 1–8. IEEE, 2024.
90. Kim Y, Xu X, McDuff D. *et al.* Health-LLM: large language models for health prediction via wearable sensor data. arXiv preprint arXiv:240106866. 2024.
91. Yang B, Jiang S, Xu L. *et al.* DrHouse: an LLM-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2024;**8**:1–29.
92. Soumma SB, Peterson D, Mehta S. *et al. Self-Supervised Learning and Opportunistic Inference for Continuous Monitoring of Freezing of Gait in Parkinson's Disease* 2025. https://arxiv.org/abs/2410.21326.
93. Imran SA, Khan MNH, Biswas S. *et al.* LLaSA: large multimodal agent for human activity analysis through wearable sensors. arXiv preprint arXiv:240614498. 2024;**3**.
94. Soumma SB, Ghasemzadeh H. GlyRAG: context-aware retrieval-augmented framework for blood glucose forecasting. 2026. https://arxiv.org/abs/2601.05353
95. Machiraju A, Farahmand E, Soumma SB. *et al.* Time-aware cross-attention for multi-modal sensor-based blood glucose forecasting. In: *2025 IEEE 21st International Conference on Body Sensor Networks (BSN)*, pp. 1–4, 2025.
96. Farahmand E, Soumma SB, Chatrudi NT. *et al.* Hybrid attention model using feature decomposition and knowledge distillation for glucose forecasting. 2025. https://arxiv.org/abs/2411.10703
97. Soumma SB, Arefeen A, Carpenter SM. *et al.* Counterfactual Modeling with fine-tuned LLMs for health intervention design and sensor data augmentation. 2026. https://arxiv.org/abs/2601.14590
98. Soumma SB, Arefeen A, Carpenter SM. *et al.* SenseCF: LLM-prompted counterfactuals for intervention and sensor data augmentation. In: *2025 IEEE 21st International Conference on Body Sensor Networks (BSN)*, pp. 1–4, 2025.
99. Chen L, Han X, Lin S. *et al.* TriMedLM: advancing three-dimensional medical image analysis with multi-modal LLM. In: *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4505–12. IEEE, 2024.
100. Dai S, Ye K, Liu G. *et al.* Zeus: zero-shot LLM instruction for union segmentation in multimodal medical imaging. *Int J Mach Learn Cybern* 2025;**16**:9067–78.
101. Lee S, Kim WJ, Chang J. *et al.* LLM-CXR: instruction-finetuned LLM for CXR image understanding and generation. arXiv preprint arXiv:230511490. 2023.
102. Kumar R, Yunbo R, Kumar J. *et al.* MID-LLM: enhancing medical image diagnostics with LLMs in a blockchain AI framework. *IEEE Internet Things J* 2025;**12**:37160–74. https://doi.org/10.1109/JIOT.2025.3583408
103. Pellegrini C, Özsoy E, Busam B. *et al.* RaDialog: a large vision-language model for radiology report generation and conversational assistance. arXiv preprint arXiv:231118681. 2023.
104. Alsinglawi B, McCarthy C, Webb S. *et al.* A lightweight large vision-language model for multimodal medical images. arXiv preprint arXiv:250405575. 2025.
105. Kumar GMK, Chadha A, Mendola JD. *et al.* MedVisionLlama: leveraging pre-trained large language model layers to enhance medical image segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1114–24, 2025.
106. Driess D, Xia F, Sajjadi MS. *et al.* PaLM-E: an embodied multimodal language model. *ICML.* 2023.
107. Singhal K, Azizi S, Tu T. *et al.* Large language models encode clinical knowledge. *Nature* 2023;**620**:172–80.
108. Nakaura T, Ito R, Ueda D. *et al.* The impact of large language models on radiology: a guide for radiologists on the latest innovations in AI. *Jpn J Radiol* 2024;**42**:685–96. https://doi.org/10.1007/s11604-024-01552-0
109. Mehandru N, Hall AK, Melnichenko O. *et al.* BioAgents: democratizing bioinformatics analysis with multi-agent systems. arXiv preprint arXiv:250106314. 2025.
110. Wang E, Schmidgall S, Jaeger PF. *et al.* TxGemma: efficient and agentic LLMs for therapeutics. arXiv preprint arXiv:250406196. 2025.
111. Mitchener L, Laurent JM, Andonian A. *et al.* BixBench: a comprehensive benchmark for LLM-based agents in computational biology. arXiv preprint arXiv:250300096. 2025.
112. Stewart I, Buehler MJ. Molecular analysis and design using generative artificial intelligence via multi-agent modeling. *Mol Syst Des Eng* 2025;**10**:314–37.
113. Singh G, Wehling L, Mulyadi AW. *et al.* Talk2Biomodels and Talk2KnowledgeGraph: AI agent-based application for prediction of patient biomarkers and reasoning over biomedical knowledge

graphs. In: *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*.

114. Xu W, Luo G, Meng W. *et al.* MRAgent: an LLM-based automated agent for causal knowledge discovery in disease via Mendelian randomization. *Brief Bioinform* 2025;**26**. https://doi.org/10.1093/bib/bbaf140
115. Philpott J, Kurjan A, Cribbs AP. FlowAgent: a modular agent-based system for automated workflow management and data interpretation. bioRxiv. 2025;2025–03.
116. DeMeo B, Nesbitt C, Miller SA. *et al.* Active learning framework leveraging transcriptomics identifies modulators of disease phenotypes. *Science* 2025;**390**. https://doi.org/10.1126/science.adi8577
117. Lu M, Ho B, Ren D. *et al.* TriageAgent: towards better multi-agents collaborations for large language model-based clinical triage. In: *Findings of the Association for Computational Linguistics: EMNLP*, Vol. **2024**, pp. 5747–64. Miami, Florida, USA, 2024.

Briefings in Bioinformatics, 2026, **27**, bbag110
https://doi.org/10.1093/bib/bbag110
Review