



LARGE LANGUAGE MODEL AGENTS FOR BIOMEDICAL RESEARCH AND DRUG DISCOVERY: FROM LITERATURE MINING TO HYPOTHESIS GENERATION

MD. Rashid Khan, Dr. Sejda Khan

Assistance Professor, Department of Computer Science, RJTE Delhi

ABSTRACT

The rapid development of large language models (LLMs), retrieval-augmented generation, multimodal artificial intelligence, and autonomous AI agents is creating new opportunities for biomedical research and pharmaceutical discovery. Conventional biomedical research requires researchers to review extensive scientific literature, identify knowledge gaps, formulate hypotheses, select relevant experimental approaches, and iteratively evaluate evidence. The rapidly expanding biomedical literature, increasing complexity of biological datasets, and multidisciplinary nature of drug discovery make these activities increasingly difficult to perform manually at scale. Large language model agents offer a potential solution by combining natural-language reasoning with information retrieval, knowledge graphs, computational tools, databases, and specialized scientific workflows. This paper examines the emerging role of LLM-based agents across the biomedical research and drug-discovery pipeline, with particular emphasis on literature mining, evidence synthesis, disease-mechanism analysis, target identification, drug repurposing, molecular design, and hypothesis generation. A conceptual agentic framework is proposed in which specialized agents perform literature retrieval, evidence extraction, knowledge synthesis, hypothesis generation, hypothesis criticism, experimental design, and validation. Recent evidence indicates that multi-agent systems can generate experimentally testable biomedical hypotheses and can support iterative scientific discovery. A 2025 study using a multi-agent LLM framework for Alzheimer's disease drug-combination discovery reported improved predictive performance over a traditional knowledge-based approach and experimentally validated one predicted combination in vitro. Recent 2026 research has further demonstrated systems that connect literature search, hypothesis generation, experimental data analysis, and hypothesis refinement within a continuous scientific workflow. Nevertheless, challenges remain concerning hallucination, evidence quality, biological validity, reproducibility, data heterogeneity, interpretability, regulatory compliance, and human oversight. The paper concludes that LLM agents should be viewed as research collaborators and evidence-synthesis systems rather than autonomous replacements for biomedical scientists. Properly designed verification, provenance, expert review, and experimental validation mechanisms are essential for responsible deployment.

KEYWORDS: Large Language Models, AI Agents, Biomedical Research, Drug Discovery, Hypothesis Generation, Literature Mining, Generative AI, Retrieval-Augmented Generation, Multi-Agent Systems, Pharmaceutical Research, Artificial Intelligence, Drug Repurposing

INTRODUCTION

Biomedical research is fundamentally dependent on the ability to integrate knowledge from a rapidly expanding scientific literature. Researchers must examine thousands of publications, identify relationships among genes, proteins, diseases, drugs, pathways, phenotypes, and clinical outcomes, and then translate these relationships into experimentally testable hypotheses. The conventional research process is highly dependent on domain expertise and human interpretation.

Drug discovery is particularly challenging because it requires the integration of chemistry, molecular biology, pharmacology, toxicology, clinical medicine, and computational science. Traditional drug discovery can require extensive time and financial investment, while many candidate compounds fail during preclinical or clinical development. Recent reviews describe LLMs as increasingly relevant to target



identification, molecular design, preclinical research, clinical-trial analysis, and other stages of drug development.

The emergence of LLMs has introduced a new computational paradigm. Unlike conventional machine-learning systems that are usually trained for a specific prediction task, LLMs can process heterogeneous scientific text and interact with researchers through natural language. When combined with retrieval systems, databases, APIs, computational tools, and autonomous planning mechanisms, LLMs can become **scientific agents** capable of performing sequences of research tasks.

The concept of an AI agent extends beyond question answering. An agent can receive a research objective, decompose the objective into subtasks, retrieve relevant evidence, use external tools, evaluate intermediate results, and generate a final research output. Multi-agent architectures further divide scientific work among specialized agents.

For example, one agent may perform literature retrieval, another may extract biological relationships, another may generate hypotheses, while a critic agent evaluates whether those hypotheses are supported by existing evidence.

Recent research demonstrates that this paradigm is moving from theoretical discussion toward practical biomedical applications. A 2026 *Nature* study introduced Robin, a multi-agent system capable of literature-guided hypothesis generation, experimental strategy development, analysis of biological experimental data, and iterative hypothesis refinement. Similarly, DeepEvidence has been developed as a multi-agent system for evidence exploration and synthesis across heterogeneous biomedical knowledge sources, including applications in drug discovery and clinical research.

Therefore, the objective of this paper is to examine how LLM agents can support biomedical research from **literature mining to hypothesis generation**, while proposing an integrated framework for evidence-grounded and experimentally verifiable AI-assisted scientific discovery.

LITERATURE REVIEW

2.1 Evolution of AI in Biomedical Research

Artificial intelligence has been applied to biomedical research for several decades, initially through expert systems, statistical models, and traditional machine-learning approaches. Deep learning subsequently enabled advances in medical imaging, genomic prediction, molecular property prediction, and drug discovery.

The introduction of foundation models and LLMs has expanded the scope of AI applications. Instead of developing a separate model for every task, researchers can use a general-purpose model and adapt it through prompting, retrieval, fine-tuning, tool use, and specialized agents.

Recent literature identifies applications of LLMs throughout the pharmaceutical development pipeline, including target identification, drug design, pharmacokinetic analysis, clinical-trial support, and scientific communication.

2.2 LLMs for Biomedical Literature Mining

Scientific literature contains enormous amounts of information about:

- genes;
- proteins;
- diseases;
- drugs;
- clinical trials;



- molecular mechanisms;
- biomarkers;
- pathways;
- adverse effects;
- therapeutic outcomes.

Traditional keyword-based literature searches may miss relationships expressed using different terminology. LLM-based systems can interpret semantic relationships between concepts and retrieve information according to research intent rather than exact keyword matching.

Recent work on OpenScholar demonstrated a retrieval-augmented language model designed to synthesize scientific literature from a corpus of approximately 45 million open-access papers. The associated benchmark included expert-written scientific queries across several disciplines, including biomedicine.

This development illustrates an important transition from simple document retrieval toward **evidence synthesis**.

2.3 Retrieval-Augmented Generation in Biomedical Research

A major weakness of standalone LLMs is that their responses may contain unsupported or hallucinated information. Retrieval-augmented generation (RAG) addresses this problem by retrieving external information before generating an answer.

In biomedical research, a RAG pipeline may operate as:

Research Question → Literature Search → Relevant Evidence → Evidence Ranking → Context Construction → LLM Synthesis → Citation

This approach enables the model to ground its responses in external scientific sources.

However, retrieval alone does not guarantee scientific validity. Retrieved papers may be outdated, contradictory, methodologically weak, or incorrectly interpreted. Consequently, biomedical RAG systems require evidence ranking, source attribution, citation verification, and domain-specific validation.

2.4 Multi-Agent Systems for Biomedical Hypothesis Generation

Multi-agent systems divide complex research activities among specialized AI agents.

One important example is the 2025 Coated-LLM framework for Alzheimer's disease drug-combination discovery. The system used multiple specialized LLM agents, including researcher, reviewer, and moderator roles, to generate and evaluate therapeutic hypotheses. The study reported higher accuracy than a traditional knowledge-based method and externally validated one predicted drug combination experimentally.

This result is important because it demonstrates that LLM agents can potentially move beyond summarization toward **hypothesis generation and scientific reasoning**.

2.5 LLM Agents in Drug Discovery

Drug discovery provides a particularly suitable environment for agentic AI because the workflow involves multiple interconnected tasks.

An agentic drug-discovery system can potentially perform:

1. disease-mechanism analysis;
2. target identification;
3. target validation;
4. literature mining;



5. molecular generation;
6. molecular-property prediction;
7. ADMET assessment;
8. drug repurposing;
9. experimental prioritization.

Recent work has proposed modular LLM agents capable of retrieving biomedical information, performing literature-grounded question answering, generating molecules, predicting molecular properties, and iteratively refining candidate molecules.

A 2026 review further describes AI agents across literature synthesis, automated protocol generation, toxicity prediction, molecular synthesis, drug repurposing, and end-to-end drug-discovery workflows.

2.6 AI for Scientific Hypothesis Generation

Scientific hypothesis generation is traditionally dependent on researchers identifying unexplained relationships or inconsistencies within existing knowledge.

LLM-based systems can assist by:

- connecting concepts across publications;
- identifying contradictory findings;
- detecting underexplored relationships;
- generating mechanistic explanations;
- proposing new drug combinations;
- identifying potential repurposing opportunities.

Recent surveys identify retrieval augmentation, knowledge graphs, simulation, causal reasoning, and multi-agent systems as important components of emerging LLM-based scientific hypothesis-generation systems.

RESEARCH GAP

Despite significant progress, several gaps remain.

Gap 1: From Literature Summarization to Discovery

Many LLM applications focus on summarizing existing research rather than generating experimentally testable hypotheses.

Gap 2: Evidence Grounding

LLMs may generate scientifically plausible but unsupported claims.

Gap 3: Hypothesis Validation

Generating a hypothesis does not establish scientific validity. Experimental or computational validation is required.

Gap 4: Multi-Agent Coordination

Different agents may produce conflicting conclusions, and mechanisms are needed to resolve disagreement.

Gap 5: Biomedical Data Integration

Biomedical knowledge exists across literature databases, genomic repositories, clinical datasets, chemical databases, and knowledge graphs.

Gap 6: Reproducibility

Scientific AI systems need detailed records of sources, prompts, tools, datasets, and intermediate decisions.



Gap 7: Human Oversight

Biomedical decisions can involve patient safety and regulatory consequences. Therefore, human researchers must remain responsible for high-impact decisions.

RESEARCH OBJECTIVES

The study has the following objectives:

1. To examine the applications of LLM agents in biomedical research.
2. To investigate their role in scientific literature mining.
3. To examine the use of LLM agents for hypothesis generation.
4. To analyze applications in drug discovery and drug repurposing.
5. To propose an integrated multi-agent biomedical research framework.
6. To identify mechanisms for evidence verification and hypothesis validation.
7. To evaluate potential benefits and limitations of agentic biomedical research.
8. To propose future directions for trustworthy autonomous scientific discovery.

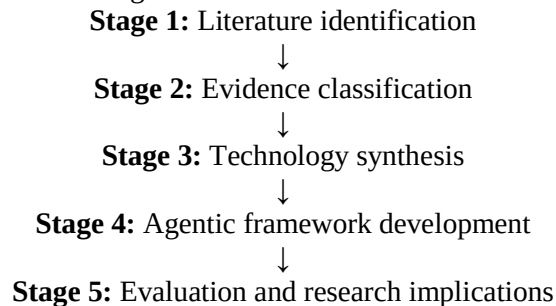
RESEARCH METHODOLOGY / MATERIALS AND METHODS

5.1 Research Design

This study adopts a **conceptual and systematic literature-based research design**.

The methodology integrates recent peer-reviewed literature and emerging scientific-agent research to develop a conceptual framework for LLM-assisted biomedical discovery.

The research process consists of five stages:



5.2 Literature Sources

Relevant scientific literature can be collected from:

- PubMed;
- PubMed Central;
- Crossref;
- Nature Portfolio;
- ScienceDirect;
- IEEE Xplore;
- SpringerLink;
- arXiv;
- Semantic Scholar.

Priority should be given to research published between **2023 and 2026**, reflecting the rapid development of LLM agents.

5.3 Proposed Agentic Biomedical Research Architecture

The proposed framework consists of eight specialized agents.

Agent 1: Research Question Agent

Receives the scientific objective and decomposes it into research questions.



Agent 2: Literature Mining Agent

Searches biomedical literature and retrieves relevant publications.

Agent 3: Evidence Extraction Agent

Extracts:

- diseases;
- genes;
- proteins;
- drugs;
- mechanisms;
- outcomes;
- experimental evidence.

Agent 4: Knowledge Graph Agent

Maps relationships between biomedical entities.

Example:

Disease → Protein → Pathway → Drug → Mechanism → Outcome

Agent 5: Hypothesis Generation Agent

Uses the synthesized evidence to generate candidate hypotheses.

Agent 6: Critic Agent

Evaluates:

- novelty;
- evidence;
- biological plausibility;
- testability;
- potential contradictions.

Agent 7: Experimental Design Agent

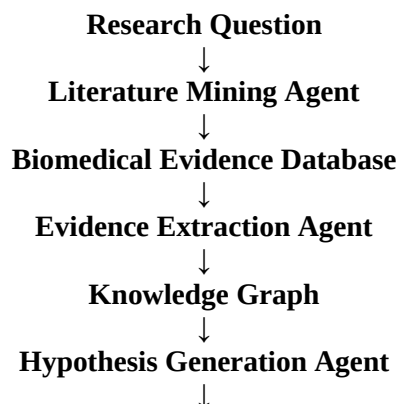
Converts selected hypotheses into computational or laboratory test plans.

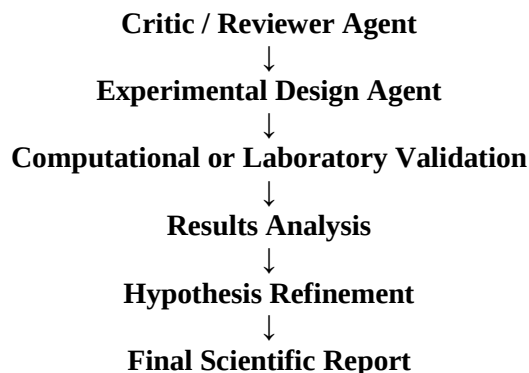
Agent 8: Validation Agent

Checks whether conclusions are supported by evidence and whether experimental results support the hypothesis.

PROPOSED WORKFLOW

The complete workflow can be represented as:





This creates an iterative research cycle rather than a one-time AI response.

RESULTS / FINDINGS

Because this paper is a conceptual research study, the results are presented as **framework-derived findings and synthesis of recent empirical evidence**, rather than fabricated experimental measurements.

7.1 Finding 1: LLM Agents Can Extend Literature Mining

Recent retrieval-augmented systems demonstrate that LLMs can search large scientific corpora and synthesize citation-backed information. OpenScholar, for example, was developed to retrieve and synthesize evidence from tens of millions of open-access papers.

This indicates that LLMs can support literature analysis at a scale that would be difficult for individual researchers.

7.2 Finding 2: Multi-Agent Architectures Improve Scientific Task Decomposition

Multi-agent systems allow research activities to be separated into specialized roles.

Instead of asking one model to perform:

Search → Analyze → Hypothesize → Criticize → Validate

different agents can perform these tasks independently.

This structure can improve traceability and make it easier to identify where an error occurred.

7.3 Finding 3: Biomedical Hypothesis Generation Is Becoming Experimentally Testable

The 2025 Coated-LLM study provides an important example. Its multi-agent architecture generated drug-combination hypotheses for Alzheimer's disease and reported both benchmark performance and experimental validation of a predicted combination.

This represents an important shift from:

AI-generated text → AI-generated scientific hypothesis → experimental validation

7.4 Finding 4: Agentic Systems Are Moving Toward Closed-Loop Discovery

The 2026 Robin system demonstrates an even broader direction by integrating literature search, hypothesis generation, experimental strategy, biological-data analysis, and iterative hypothesis refinement.

This suggests the development of a closed-loop scientific workflow:

Knowledge → Hypothesis → Experiment → Data → New Knowledge
Knowledge → Hypothesis → Experiment → Data → New Knowledge

7.5 Finding 5: Drug Discovery Is a High-Potential Application

The literature indicates that LLM agents can potentially support several stages of drug discovery, including molecular generation, property prediction, literature analysis, repurposing, toxicity assessment, and experimental prioritization.

However, computational success should not be interpreted as proof of clinical effectiveness.



7.6 Finding 6: Evidence Provenance Is Essential

Recent biomedical research-agent systems increasingly emphasize evidence tracing and knowledge graphs. DeepEvidence, for example, constructs an evidence graph to support tracking, attribution, and validation of the research process.

This suggests that future biomedical agents should maintain provenance for every major scientific claim.

DISCUSSION

The findings indicate that LLM agents are changing the role of AI in biomedical research.

Traditional AI systems typically perform predefined prediction tasks. Agentic systems introduce a more flexible architecture in which AI can coordinate multiple research activities.

The most significant development is the transition from **information retrieval to scientific reasoning**.

A conventional search engine answers:

"Which papers discuss this topic?"

An LLM research agent can potentially answer:

"What relationships appear consistently across these studies?"

A more advanced multi-agent system can ask:

"What scientifically testable hypothesis emerges from these relationships?"

The next step is:

"What experiment could test this hypothesis?"

Finally:

"What does the experimental evidence imply, and how should the hypothesis be modified?"

This creates a continuous scientific workflow.

However, several risks remain.

8.1 Hallucination

LLMs can produce plausible but incorrect scientific claims. In biomedical research, such errors can be particularly serious.

8.2 False Novelty

An AI-generated hypothesis may appear novel but already exist in an obscure publication. Comprehensive literature retrieval is therefore essential.

8.3 Biological Plausibility

A statistically or linguistically plausible hypothesis may not be biologically meaningful.

8.4 Experimental Translation

A computational prediction does not guarantee experimental or clinical success.

8.5 Data Bias

Biomedical literature may overrepresent successful experiments and established research areas while underrepresenting negative results.

8.6 Human Oversight

AI-generated hypotheses should be reviewed by domain experts before laboratory implementation.

RECOMMENDATIONS / IMPLICATIONS

9.1 Implications for Biomedical Researchers

Researchers can use LLM agents as research collaborators for:

- systematic literature exploration;
- evidence synthesis;



- research-gap identification;
- hypothesis brainstorming;
- protocol development;
- data-analysis assistance.

However, researchers should independently verify important claims.

9.2 Implications for Pharmaceutical Companies

Pharmaceutical organizations can integrate agentic AI into early-stage discovery workflows for:

- target prioritization;
- drug repurposing;
- compound generation;
- literature surveillance;
- biomarker discovery;
- experimental prioritization.

9.3 Implications for Academic Institutions

Universities can establish multidisciplinary research environments combining:

AI + Biology + Chemistry + Pharmacology + Data Science

Such collaboration can accelerate development of scientifically grounded AI systems.

9.4 Implications for AI Developers

Developers should prioritize:

- evidence grounding;
- provenance tracking;
- uncertainty estimation;
- citation verification;
- tool-use security;
- reproducibility;
- expert feedback.

9.5 Regulatory Implications

AI-generated biomedical findings should be clearly distinguished from experimentally validated evidence. Regulatory and clinical decisions should continue to rely on appropriate validated evidence and expert oversight.

LIMITATIONS

This study has several limitations.

First, the proposed framework is primarily conceptual and does not present a newly collected biomedical dataset or prospective laboratory experiment.

Second, the performance of LLM agents depends heavily on the underlying language model, retrieval system, biomedical databases, and quality of scientific literature.

Third, scientific novelty is difficult to quantify automatically.

Fourth, agentic systems may propagate errors across multiple agents if all agents rely on the same incorrect source.

Fifth, computational evaluation cannot replace biological experimentation.

Finally, the field is evolving rapidly, meaning that agent architectures, models, and benchmarks may change substantially over short periods.



FUTURE RESEARCH

Future research should focus on experimentally validating LLM-agent architectures in specific biomedical domains.

11.1 Autonomous Drug Repurposing

Agents could identify existing drugs that may target previously unexplored diseases.

11.2 Multimodal Biomedical Agents

Future systems should combine:

- text;
- molecular structures;
- genomic sequences;
- protein structures;
- medical images;
- clinical records.

11.3 Knowledge-Graph-Enhanced Agents

Knowledge graphs can provide structured relationships among biomedical entities.

11.4 Autonomous Design-Make-Test-Analyze Systems

Future systems could connect computational molecular design with automated laboratory platforms.

11.5 Causal Hypothesis Generation

Future research should move beyond correlation toward causal reasoning.

11.6 Experimental Validation

AI-generated hypotheses should increasingly be evaluated through controlled laboratory experiments.

11.7 Human–AI Scientific Collaboration

Future studies should examine whether AI agents improve the productivity and quality of research teams rather than merely measuring model performance.

CONCLUSION

Large language model agents represent an emerging paradigm for biomedical research and drug discovery. Their ability to combine natural-language reasoning, literature retrieval, knowledge synthesis, external tools, and specialized agent roles creates opportunities to automate portions of the scientific discovery process.

The literature reviewed in this paper indicates that the field is moving beyond simple question answering toward evidence synthesis, hypothesis generation, experimental planning, and iterative scientific discovery. Recent systems such as multi-agent biomedical hypothesis-generation frameworks, retrieval-augmented scientific research systems, DeepEvidence, and Robin demonstrate different components of this emerging paradigm.

The proposed framework integrates these capabilities into a research pipeline consisting of literature mining, evidence extraction, knowledge-graph construction, hypothesis generation, critical evaluation, experimental design, and validation.



Nevertheless, LLM agents should not currently be regarded as autonomous replacements for biomedical researchers. Their outputs remain vulnerable to hallucination, incomplete evidence, false novelty, data bias, and insufficient biological validation.

The most appropriate near-term model is therefore **human–AI collaborative scientific discovery**, where LLM agents accelerate information processing and hypothesis generation while researchers retain responsibility for scientific judgment, experimental validation, ethical considerations, and interpretation. The future of AI-enabled drug discovery is consequently unlikely to depend on language models alone. Instead, progress will increasingly involve the integration of **LLMs, retrieval systems, biomedical knowledge graphs, molecular models, multimodal AI, autonomous agents, computational tools, and experimental laboratories** into interconnected scientific discovery ecosystems.

REFERENCES

1. Boiko, D. A., MacKnight, R., & Gomes, G. (2023). Emergent autonomous scientific research capabilities of large language models. *Nature Communications*, 14, 5980. <https://doi.org/10.1038/s41467-023-44591-7>
2. Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., & Schwaller, P. (2024). Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6, 525–535. <https://doi.org/10.1038/s42256-024-00832-8>
3. Das, S., Ferdaus, M. M., & Dam, T. (2026). AI agents in drug discovery: A review of evolution, applications, and future directions. *Computer Methods and Programs in Biomedicine*, 285, 109539. <https://doi.org/10.1016/j.cmpb.2026.109539>
4. Guo, T., Chen, X., Wang, Y., Chang, Y., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16), 17919–17927. <https://doi.org/10.1609/aaai.v38i16.29796>
5. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55. <https://doi.org/10.1145/3703155>
6. Lu, J., Choi, K., Ereemeev, M., Gobburu, J., Goswami, S., Liu, Q., Mo, G., Musante, C. J., & Shahin, M. H. (2025). Large language models and their applications in drug discovery and development: A primer. *Clinical and Translational Science*, 18(4), e70205. <https://doi.org/10.1111/cts.70205>
7. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2024). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.
8. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., He, X., & others. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18, 186345. <https://doi.org/10.1007/s11704-024-40231-1>
9. Wang, Z., Danek, B., Yang, Z., Chen, Z., & Sun, J. (2026). Making large language models reliable data science programming copilots for biomedical research. *Nature Biomedical Engineering*, 10, 1732–1746. <https://doi.org/10.1038/s41551-025-01587-2>
10. Xu, Q., Soto, C., Shahnawaz, M., Liu, X., Jiang, X., & Kim, Y. (2025). Multi agent large language models for biomedical hypothesis generation in drug combination discovery. *iScience*, 28(12), 113984. <https://doi.org/10.1016/j.isci.2025.113984>
11. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2210.03629>



12. Zheng, Y., Koh, H. Y., Ju, J., Yang, M., May, L. T., Webb, G. I., Li, L., Pan, S., & Church, G. (2025). Large language models for drug discovery and development. *Patterns*, 6(10), 101346. <https://doi.org/10.1016/j.patter.2025.101346>
13. Zheng, Y., et al. . A collaborative large language model for drug analysis. *Nature Biomedical Engineering*, 10, 870–881. <https://doi.org/10.1038/s41551-025-01471-z>
14. DISRUPT-DS Industry Roundtable. (2026). Generative AI in pharmaceutical R&D: From large language models to AI agents to regulation. *Drug Discovery Today*, 31(1), 104593. <https://doi.org/10.1016/j.drudis.2025.104593>
15. Wei, J., Yang, Y., Zhang, X., Chen, Y., Zhuang, X., Gao, Z., Zhou, D., Wang, G., Gao, Z., Cao, J., Qiu, Z., He, X., Zhang, Q., You, C., Zheng, S., Ding, N., Ouyang, W., Dong, N., Cheng, Y., Sun, S., Bai, L., & Zhou, B. (2025). From AI for science to agentic science: A survey on autonomous scientific discovery. *arXiv*. <https://doi.org/10.48550/arXiv.2508.14111>