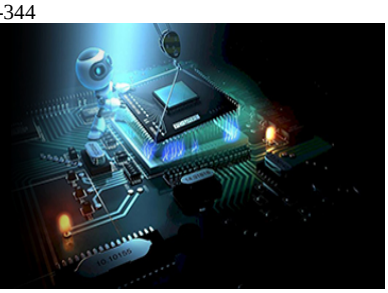
International Journal of Engineering in Computer Science



E-ISSN: 2663-3590
P-ISSN: 2663-3582
Impact Factor (RJIF): 5.52
www.computersciencejournals.com/ijecs
IJECS 2025; 7(2): 339-344
Received: 24-10-2025
Accepted: 29-11-2025

Dheeraj Sonkhla
Department of Computer Science, Himachal Pradesh University Regional Centre, Dharamshala, Himachal Pradesh, India

Corresponding Author:
Dheeraj Sonkhla
Department of Computer Science, Himachal Pradesh University Regional Centre, Dharamshala, Himachal Pradesh, India

A review of large language models for pharmaceutical advice: Techniques, challenges, and applications

Dheeraj Sonkhla

DOI: https://www.doi.org/10.33545/26633582.2025.v7.i2d.237

Abstract
The use of Large Language Models (LLMs) has significantly transformed the digital healthcare. Pharmacy is a distinguished part of healthcare and recent researches in this domain have increased tremendously. This review article is focus on current advancements, techniques and data requirements of LLMs when dealing with pharmaceutical field. The paper analyzes the core techniques including Chain-of-Thought (CoT) prompting for structured reasoning, domain-specific fine-tuning, and Knowledge Graph integration to ensure interpretability. The Retrieval-Augmented Generation (RAG) are discussed in detail as it makes the LLM responses to mitigate hallucinations and improve accuracy. Despite these advancements, the review highlights critical limitations in current architectures. LLMs exhibit significant fragility to input noise and struggle with complex clinical guideline adherence. Notably, general-purpose models like GPT-4 demonstrated a 71% failure rate in detecting potential drug-drug interactions (pDDIs) compared to standard software, posing serious safety risks. The study concludes that while LLMs offer unprecedented opportunities for efficiency and information synthesis, they cannot yet function as autonomous agents. Safe implementation requires hybrid human-AI workflows, robust adversarial defenses, and harmonized regulatory frameworks to validate performance in high-stakes pharmaceutical environments.

Keywords: Large language model, knowledge graph, generative artificial intelligence, natural language processing

1. Introduction
In recent times, Generative Artificial Intelligence (GAI), especially Large Language Models (LLMs), has revolutionized artificial intelligence and gained significant attention. LLMs, such as OpenAI's GPT-4, Google's Gemini, and Meta's LLaMA, have demonstrated remarkable proficiency in generating and understanding natural language, thereby transforming various domains [1]. Initially focused on Natural Language Processing (NLP) and comprehension, these models now have evolved to support multimodal capabilities such as decision-making, content creation, and problem-solving [2].

LLMs have various applications across different domains. In education, they assist in creating personalized learning experiences and tutoring systems [3]. Their ability to process and interpret large volumes of data makes them invaluable in addressing intricate challenges across these fields. In healthcare and pharmaceuticals, the potential of LLMs is particularly promising. The healthcare sector is known for its data-intensive nature. The health professionals often overwhelmed by the sheer volume of medical information, research, and patient data. LLMs can alleviate this burden by assisting in various tasks, such as diagnosing diseases, providing treatment recommendations, and retrieving up-to-date drug information [1]. One such application includes a model developed to detect and provide real-time advice for Parkinson's disease [4]. There are other examples like, LLM designed to assist with the Medical Licensing Exam [5], and a medication advice model tailored to offer comprehensive guidance on drug interactions and contraindications [6].

Despite their immense capabilities, existing LLMs in the healthcare domain face limitations, especially in providing precise medication guidance and identifying untoward drug reactions [6]. This review article aims to explore the current advancements in training and employing customised LLMs in healthcare. The purpose is to delve into the methodologies, data requirements, and evaluation metrics necessary to develop robust models capable of providing accurate and comprehensive medication guidance specifically for pharmaceutical

advisory roles. By exploring and analysing the challenges and potential solutions, this study aspires to contribute to the integration of generative AI into the pharmaceutical industry.

2. Literature Review

The applications of LLMs in the healthcare are accelerating in the recent times. The capability of LLMs to deal with the diverse data helps greatly in introducing various solutions to the medical issues. This section provides a review of some recent works done in the healthcare by incorporating the generative AI and LLMs.

2.1 Parkinson's Disease Management

The research done by [4] includes the use of Internet of Things (IoT) and LLMs for developing a personalised butler for the patients with the Parkinson's Disease (PD). The research developed an automatic wearable system which provides real-time PD diagnosis, monitoring, and personalised recommendations. The device offers affordable and portable accessibility to the users.

2.2 Lung Cancer Detection

Lung cancer continues to be a major challenge in global health. The study [7] introduces a deep learning framework that uses the capabilities of LLMs alongside medical imaging data to enhance the accuracy of lung cancer detection. The proposed system integrates the patient-reported symptoms, medical images, and doctors' prescriptions into a comprehensive dataset. Deep learning techniques including Convolutional Neural Networks (CNNs) are trained with this dataset for image analysis and LLMs for textual data interpretation. The results of the research reveals that the model outperforms existing systems in lung cancer detection efficiency.

2.3 LLM tuning for Medical Guidance

Nowadays, most of the research in LLMs is being done in the field of recommendations or guidance such as movie, anime, or medical recommendations [8][2]. The authors [6] introduced their own LLM named ShennongGPT, designed specially for the medical guidance. The major restriction of the work is that it only works for the Chinese language and no other language.

2.4 GAI for Dental Licensing Examinations

To know the efficiency of LLMs, research has been carried out by [5]. The study includes the comparison of two LLM versions, ChatGPT 3.5 and ChatGPT 4.0. Research includes the accuracy comparison of GenAI in answering the questions from Dental licensing exams for UK and US. The ChatGPT 4.0 outperforms the ChatGPT 3.5 by approx. 15%. The major drawback of this study is that, these observations may not hold true for the scenarios.

To have the better understanding of LLMs and their applications in the healthcare, it is necessary to review articles with the similar research area. Table 1 shows the research papers in the field of healthcare which have employed the LLMs.

Table 1: A summary of various research papers using LLMs

References	Focus Area	Summary	Challenges and Limitations
[2]	Healthcare	• Discussed the various aspects of GAI in Healthcare like applications and challenges	Security and privacy issue Training data requirement Biasness
[6]	Healthcare	• Introduced a new LLM model, ShennongGPT for medication guidance and untoward drug reaction	Trained in Chinese language Size of data utilised is confined
[5]	Healthcare and education	• Performed the LLM models to analyse their performance in medical licensing exams	Limited research work
[7]	Healthcare	• Integrated the LLMs and deep learning models for advanced lung cancer detection • Diverse data modalities are used for the better performance	Data heterogeneity requires advanced preprocessing strategies Ethical concerns Biasness due to limited data
[9]	Healthcare	• Discussed the various applications and challenges of ChatGPT in Healthcare	Lack of contextual and factual knowledge
[1]	Healthcare	• Reviewed the potential applications of ChatGPT and other LLMs in Healthcare	Misinformation and Biasness risk in ChatGPT
[4]	Healthcare	• Use of AI in Parkinson's disease management • Real time monitoring, diagnosis and recommendations for the patients	enhancing the usability and effectiveness of wearable devices and smartphone tools to ensure better patient compliance
[10]	Healthcare	• Reviewed the multi—modal approach for healthcare • Topics covered are medical imaging and GAI based LLMs	• Baises in training data • Patient data privacy • NLP challenges

3. Core Techniques and Methodologies

3.1 Prompt Engineering and Chain-of-Thought Reasoning

Prompt engineering has emerged as a critical technique for optimizing LLM performance in pharmaceutical applications. Among documented approaches, Chain-of-Thought (CoT) prompting demonstrates superior reasoning performance compared to zero-shot and few-shot methods, with self-consistency (an ensemble-based variant of CoT) consistently outperforming standard CoT across diverse medical tasks [11].

Medical-specific CoT approaches, such as Diagnosis Reasoning Chain-of-Thought (DR-CoT), enforce clinical protocols by requiring explicit summarization of evidence, iterative differential diagnosis ranking, and systematic inquiry generation. These structured reasoning scaffolds mirror established clinical workflows and improve out-of-domain generalization by 18% in medical applications. The effectiveness of CoT prompting reflects its ability to

decompose complex pharmaceutical reasoning into interpretable intermediate steps that align with clinical decision-making processes [12].

Few-shot prompting, where models are provided with examples of correct responses, shows variable effectiveness depending on task complexity and domain specificity. Recent developments emphasize the importance of clinician-approved reasoning pathways, with reward models and majority voting mechanisms enhancing both accuracy and transparency in medical outputs [13].

3.2 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation has proven to be one of the most impactful techniques for improving LLM accuracy in pharmaceutical contexts. RAG systems combine language models with external knowledge retrieval mechanisms, enabling LLMs to ground responses in authoritative pharmaceutical and biomedical literature [14].

Performance improvements are substantial: on PubMedQA (biomedical question-answering), GPT-4 without retrieval achieved 57.9% accuracy, while RAG-enabled systems achieved 86.3% a 28.4 percentage point improvement. Similarly, in BioASQ benchmarks, RAG elevated GPT-3.5 accuracy from approximately 74% to 90%. The MIRAGE benchmark (Medical Information Retrieval-Augmented Generation Evaluation), encompassing 7,663 questions across five biomedical datasets, demonstrated that RAG improved LLM accuracy by up to 18% over standard chain-of-thought prompting [14].

RAG is particularly valuable for pharmaceutical advice because it addresses the hallucination problem by explicitly grounding responses in retrieved documents. When querying updated clinical guidelines, drug interaction databases, or recent clinical trials, RAG systems can synthesize current evidence rather than relying solely on training data. Smaller models like LLaMA-2 70B with RAG approach the performance of specialized biomedical models when equipped with domain-specific knowledge repositories [14].

3.3 Fine-Tuning and Domain-Specific Models

Domain-specific LLMs trained on pharmaceutical and biomedical literature have shown substantial performance advantages over general-purpose models. BioformerTM, a compact BERT variant trained on 33 million PubMed abstracts and 1 million PMC articles, achieves 99% of the performance of PubMedBERT while using only 40% of the parameters, with 2-3x faster inference. This efficiency is critical for real-world pharmaceutical applications where latency and computational costs directly impact deployment feasibility [15].

BioBERT, SciBERT, and ClinicalBERT have been developed specifically for biomedical NLP tasks including named entity recognition, relation extraction, and clinical text classification. Fine-tuning these models on specialized pharmaceutical datasets dramatically improves performance on downstream tasks. For adverse drug event (ADE) detection, fine-tuned LLMs achieved 85-86% accuracy with AUC of 87%, substantially surpassing traditional machine learning approaches [16][17].

Specialized therapeutics-focused models such as Tx-LLM (developed by Google) span the entire drug discovery pipeline, from target discovery through clinical trial approval strategy. These models demonstrate the feasibility of creating generalist models with specialized pharmaceutical training [18].

3.4 Knowledge Graph Integration

Knowledge graphs (KGs) provide structured representations of pharmaceutical relationships drug-drug interactions, drug-target interactions, adverse effects, and disease associations that can be integrated with LLMs to improve accuracy and interpretability. KnowDDI represents a state-of-the-art approach that combines graph neural networks with knowledge graphs to predict drug-drug interactions while providing interpretable reasoning paths. The method learns knowledge subgraphs specific to each drug pair, identifying explaining paths through known interactions and similar drug relationships [19].

This integration addresses a critical limitation of standalone LLMs: the ability to reason over structured pharmaceutical knowledge with guaranteed fidelity. Knowledge graph-augmented approaches have shown particular promise for detecting previously unknown drug interactions and for ensuring recommendations align with established pharmaceutical relationships [19].

Data Collection

To transform a general-purpose Large Language Model into a reliable pharmaceutical assistant, it must be grounded in diverse, authoritative data ecosystems. Reliance on pre-training alone carries the risk of hallucination; therefore, integrating structured data ranging from regulatory formularies to real-world patient records is essential for clinical accuracy and safety. The following table summarizes eight foundational categories of pharmacological data, detailing their specific sources, technical applications, and the critical value they add to creating safe, evidence-based AI systems in healthcare.

Table 2: Different sources of pharmacy data

Data Type	Description	Key Sources	LLM Application & Functionality	Impact
National Formulary	Official list of approved medicines, formulations, dosages, and regulatory status.	• British National Formulary (BNF) • USP • Indian Pharmacopoeia	Clinical Decision Support Tool: Retrieves drug info, suggests personalized dosing, checks interactions, and predicts efficacy.	Enhanced safety, reduced adverse reactions, and streamlined clinical workflows.
Medicine APIs	Biologically active components responsible for therapeutic effects (chemical properties, MOA).	• DrugBank • PubChem • KEGGDrug Database	Drug Interaction Prediction Tool: Predicts interactions based on chemical properties, issues safety alerts, and suggests alternatives.	Minimizes risk of adverse interactions and supports informed prescribing.
Monographs	Comprehensive deep-dives on individual drugs (chemistry, contraindications, adverse effects).	• Martindale • AHFS Drug Information	Drug Information Retrieval Tool: Provides detailed Q&A, comprehensive data access, and general decision support.	Increases access to detailed knowledge and ensures safer prescribing practices.

		• EMA Monographs		
Pharmacovigilance Data	Monitoring of medicine safety, specifically Adverse Drug Reactions (ADRs).	• FDA FAERS • WHO VigiBase • EudraVigilance	Adverse Reaction Prediction Tool: Detects potential ADRs from historical data, issues emerging safety alerts, and automates regulatory reporting.	Improves detection of safety signals and ensures regulatory compliance.
Clinical Trials Data	Data from study designs, results, and adverse events in controlled trials.	• ClinicalTrials.gov • EudraCT • ICTRP	Efficacy Prediction Tool: Synthesizes trial data to predict treatment outcomes and assists in evidence-based selection.	Supports evidence-based medicine and guides future clinical research.
Electronic Health Records (EHRs)	Real-world digital patient history (demographics, diagnoses, treatment plans).	• Hospital/Clinic EHR systems • Health Information Exchanges (HIEs)	Personalized Medicine Tool: Analyzes patient history to identify patterns, predict response, and tailor recommendations.	Enhances treatment personalization and supports data-driven clinical decisions.
Pharmacokinetics (PK) & Pharmacodynamics (PD)	Data on how the body affects the drug (PK) and how the drug affects the body (PD).	• Clinical pharmacology studies • Drug labels/submissions	Dosing Optimization Tool: Models drug behavior (absorption/metabolism) to optimize dosing regimens and monitor therapeutic levels.	Ensures therapeutic efficacy while minimizing toxicity risks.
Genomic Data	Genetic makeup information influencing drug susceptibility and metabolism.	• GWAS • PharmGKB • Personalized medicine initiatives	Pharmacogenomics Tool: Analyzes genetic variants (e.g., CYP enzymes) to predict individual drug response and assess risk.	High-level personalization and mitigation of genetic-specific adverse risks.

5. Challenges and Considerations

The report paints a picture of Large Language Models as powerful but brittle engines of knowledge. While LLMs demonstrate "surface-level" mastery often outperforming pharmacy students in rote memorization and fact recall they exhibit significant degradation when asked to apply that knowledge to complex, real-world clinical reasoning. This "application gap" is most visible in their inability to synthesize multiple clinical guidelines simultaneously and their dangerous underperformance in detecting drug-drug interactions, where they missed over 70% of alerts caught by standard systems. Furthermore, the models are technically fragile; minor input errors (like typos) can corrupt their outputs, and they remain vulnerable to adversarial security attacks. Consequently, the report concludes that these issues are inherent to the current architecture and cannot be solved simply by feeding the models more data; instead, external verification systems (like RAG and Knowledge Graphs) are strictly necessary.

Table 3: Challenges and Implications

Challenge	Key Findings & Evidence	Implication for Pharma
Hallucination & Inaccuracy	• Fragility: Non-clinical noise (typos, whitespace) can drop treatment accuracy by 7-9%. • Inherent Nature: Hallucination is mathematically inevitable in LLMs and cannot be solved by training alone. • Exception: Hallucinations may aid creative molecular discovery, but are dangerous for clinical advice.	Reliance on architectural safeguards (RAG, Knowledge Graphs) is mandatory; LLMs cannot be trusted as standalone agents for safety-critical advice.
Knowledge Application Gap	• Recall vs. Reasoning: LLMs excel at fact recall (93% s. 84% for students) but fail at applying that knowledge to clinical scenarios (68% vs. 80% for students).	LLMs function well as encyclopedias but poorly as clinicians, struggling to synthesize facts into actionable care plans.
Guideline Adherence Failures	• Complexity Struggle: Performance drops significantly when multiple guidelines must be considered simultaneously. • Inconsistency: The AMEGA benchmark shows highly variable performance across specialties (e.g., better in psychiatry, worse in complex multi-condition scenarios).	LLMs cannot yet reliably navigate the nuance of complex, multi-step clinical protocols required for standard of care.
DDI Detection Failure	• Severe Underperformance: GPT-4 identified only 80 potential drug-drug interactions (pDDIs) compared to 280 identified by standard software (a 71% miss rate).	This represents a critical patient safety risk; LLMs are currently inferior to standard, rule-based databases for interaction checking.
Security Vulnerabilities	• Adversarial Attacks: Models are susceptible to "poisoned" data and prompt injections that force harmful recommendations. • Regression: Newer models (e.g., LLaMA-3.3) often show increased vulnerability compared to predecessors.	Security is not improving with model scaling; specific defensive architectures are required to prevent malicious manipulation.

6. Future Directions

6.1 Specialized Pharmaceutical Language Models

Development of foundation models specifically pre-trained on pharmaceutical knowledge (drug interaction databases, prescribing guidelines, adverse event databases, clinical trial data) represents a promising direction. These models could achieve higher performance on pharmaceutical tasks while requiring less explicit knowledge augmentation.

6.2 Multimodal Integration

Integration of LLMs with structured pharmaceutical data (molecular structures, genomic data, imaging), electronic

health records, and temporal patient information could improve reasoning about complex clinical scenarios. Few existing systems leverage multimodal information effectively, representing an underexplored opportunity.

6.2 Regulatory Framework Development
Harmonized regulatory approaches to LLM validation, post-market surveillance, and accountability mechanisms will be critical for broader pharmaceutical adoption. Collaborative development involving FDA, manufacturers, and healthcare systems could accelerate responsible deployment.

6.3 Robust Adversarial Defenses
Developing architectural solutions to improve robustness against adversarial attacks—distinct from model scaling—represents an essential direction. Ensemble methods, input validation, and anomaly detection approaches may improve pharmaceutical AI security.

7. Conclusion
Large Language Models present unprecedented opportunities for enhancing pharmaceutical decision support, drug discovery, and clinical outcomes. However, their integration into pharmaceutical practice remains constrained by inherent limitations in knowledge application, hallucination, clinical guideline adherence, and regulatory uncertainty. Current evidence supports LLM deployment in well-constrained contexts where human expertise remains central: assisting with literature review and information synthesis, supporting but not replacing clinical judgment, and enhancing efficiency through automation of knowledge-based tasks.

Safe and effective pharmaceutical LLM applications require hybrid human-AI workflows incorporating retrieval-augmented generation, knowledge graph integration, explicit verification mechanisms, robust explainability, and rigorous adherence to regulatory frameworks. Future development should prioritize domain-specific models trained on pharmaceutical knowledge, multimodal integration with clinical data, improved adversarial robustness, and collaborative regulatory frameworks that enable innovation while protecting patient safety.

The pharmaceutical industry is well-positioned to leverage LLM capabilities responsibly if implementations prioritize transparency, verification, human oversight, and continuous monitoring rather than pursuing autonomous AI decision-making in this high-stakes domain.

References

1. Ali H, Qadir J, Alam T, Househ M, Shah Z. ChatGPT and large language models in healthcare: opportunities and risks. Proceedings of the 2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings). 2023. doi:10.1109/AIBThings58340.2023.10291020.
2. Kuzlu M, Xiao Z, Sarp S, Catak FO, Gurler N, Guler O. The rise of generative artificial intelligence in healthcare. Proceedings of the 12th Mediterranean Conference on Embedded Computing (MECO). 2023. doi:10.1109/MECO58584.2023.10155107.
3. Ratten V, Jones P. Generative artificial intelligence (ChatGPT): implications for management educators. International Journal of Management Education. 2023;21(3):100857. doi:10.1016/j.ijme.2023.100857.
4. Cardenas L, Parajes K, Zhu M, Zhai S. AutoHealth: advanced LLM-empowered wearable personalized medical butler for Parkinson's disease management. Proceedings of the 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC). 2024;375-379. doi:10.1109/CCWC60891.2024.10427622.
5. Chau RCW, Thu KM, Yu OY, Hsung RTC, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. International Dental Journal. 2024. doi:10.1016/j.identj.2023.12.007.
6. Dou Y, Zhao X, Zou H, Xiao J, Xi P, Peng S. ShennongGPT: a tuning Chinese LLM for medication guidance. Proceedings of the 1st IEEE International Conference on Medical Artificial Intelligence (MedAI). 2023;67-72. doi:10.1109/MedAI59581.2023.00017.
7. Vallabhaneni GV, Rahul YS, Kumari KS. Improved lung cancer detection through use of large language systems with graphical attributes. Proceedings of the International Conference on Computational and Communication Technologies. 2024;1849-1854. doi:10.1109/IC2PCT60090.2024.10486290.
8. Agarwal A, Sharma S. LLANIME: large language models for anime recommendations. Proceedings of the International Conference on Developments in ESystems Engineering (DeSE). 2023;870-875. doi:10.1109/DeSE60595.2023.10468757.
9. Mosaiyebzadeh F, Pouriyeh S, Parizi RM, Han M, Dehbozorgi N, Dorodchi M, Batista DM. Empowering healthcare professionals and patients with ChatGPT: applications and challenges. Proceedings of the International Conference on Computer Science and Education. 2024;1-7. doi:10.1109/CSCE60160.2023.00233.
10. Sathe N, Deodhe V, Sharma Y, Shinde A. A comprehensive review of AI in healthcare: exploring neural networks in medical imaging, LLM-based interactive response systems, NLP-based EHR systems, ethics, and beyond. Proceedings of the 2023 International Conference on Advanced Computing and Communication Technologies (ICACCTech). 2023;633-640. doi:10.1109/ICACCTech61146.2023.00108.
11. Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C. Prompt engineering paradigms for medical applications: scoping review. Journal of Medical Internet Research. 2024. doi:10.2196/60501.
12. Emergent Mind. Medical chain-of-thought reasoning. Emergent Mind. 2025 Oct 7. Accessed 2025 Dec 15. Available from: https://www.emergentmind.com/topics/medical-chain-of-thought-reasoning
13. Nachane SS, *et al*. Few-shot chain-of-thought driven reasoning to prompt LLMs for open-ended medical question answering. Findings of the Association for Computational Linguistics: EMNLP 2024. Miami (FL); 2024 Nov. p. 542-573. doi:10.18653/v1/2024.findings-emnlp.31.
14. Laurent A. Performance of retrieval-augmented generation (RAG) on pharmaceutical documents. Intuition Labs. 2025 Dec 14. Accessed 2025 Dec 15. Available from: https://intuitionlabs.ai/articles/rag-performance-pharmaceutical-documents

15. Fang L, Chen Q, Wei CH, Lu Z, Wang K. Bioformer: an efficient transformer language model for biomedical text mining. arXiv preprint. 2023 Feb 3:arXiv:2302.01588v1. PMID:36945685; PMCID:PMC10029052.
16. Westat. Fine-tuning LLMs to improve adverse drug event detection and reporting. Westat. 2025 Apr 7. Accessed 2025 Dec 15. Available from: https://www.westat.com/insights/llms-adverse-drug-event-detection-reporting/
17. Sabbir MAI. Data-driven healthcare: exploring biomedical text mining through NLP models. Scientific Research Journal of Science, Engineering and Technology. 2024;2(2):47.
18. Lu J, Choi K, Eremeev M, Gobburu J, Goswami S, Liu Q, Mo G, Musante CJ, Shahin MH. Large language models and their applications in drug discovery and development: a primer. Clinical and Translational Science. 2025 Apr;18(4):e70205. doi:10.1111/cts.70205. PMID:40208836; PMCID:PMC11984503.
19. Wang Y, Yang Z, Yao Q. Accurate and interpretable drug-drug interaction prediction enabled by knowledge subgraph learning. Communications Medicine. 2024;4:59. doi:10.1038/s43856-024-00486-y.