, , pp. 1–8

doi:

Prompt-to-Pill: Multi-Agent Drug Discovery and Clinical Simulation Pipeline

Ivana Vichentijevikj,[1] Kostadin Mishev[1] and Monika Simjanoska Misheva[1]

[1]Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Rugjer Boshkovikj 16, P.O. 393, 1000, Skopje, North Macedonia

Abstract

This study presents a comprehensive, modular framework for AI-driven drug discovery and clinical trial simulation, spanning from target identification to virtual patient recruitment. Synthesized from a systematic analysis of 51 LLM-based systems, the proposed *Prompt-to-Pill**architecture and corresponding implementation leverages a multi-agent system divided into preclinical and clinical phases, coordinated by a central *Orchestrator*. Each phase comprises specialized large language model (LLM) agents responsible for tasks such as molecular generation, toxicity screening, docking simulation, trial design, and patient matching. To demonstrate the full pipeline in practice, the well-characterized target Dipeptidyl Peptidase 4 (DPP4) was selected as a representative use case. The process begins with generative molecule creation and proceeds through ADMET evaluation, structure-based docking, and lead optimization. Clinical-phase agents then simulate trial generation, patient eligibility screening using EHRs, and predict trial outcomes. By tightly integrating generative, predictive, and retrieval-based LLM components, this architecture bridges preclinical discovery with virtual clinical development, offering an end-to-end demonstration of how LLM-based agents can operationalize the entire drug development workflow in silico.

Key words: Large Language Models; LLMs; Drug Discovery; Preclinical Phase; Clinical Phase; Multi-Agent Systems; Prompt-to-Pill; ChatMED

Introduction

The ability of LLMs to learn from massive datasets and adapt to diverse inputs provides unprecedented capabilities that surpass traditional methods. Their potential to revolutionize drug development extends beyond scientific advancements, impacting the economic landscape by accelerating decision-making and reducing experimental overhead (1), as evidenced by applications ranging from generating novel molecular structures (2; 3) to predicting pharmacokinetics and toxicity (4; 5), simulating clinical trials (6), and optimizing patient-trial matching (7; 8).

The integration of LLMs into drug discovery pipelines has gained notable traction, especially across preclinical phases.

Gao et al. proposed a domain-guided multi-agent system for reliable drug-target interaction (DTI) prediction, using a debate-based ensemble of LLMs. The framework partitions the DTI task into protein sequence understanding, drug structure analysis, and binding inference, handled by dedicated agents. Evaluation was conducted on the BindingDB dataset, showing improvements in both accuracy and prediction consistency compared to single-LLM baselines. The system integrates *GPT-4o*, *LLaMA-3*, and *GLM-4-Plus* (9).

Lee et al. developed *CLADD*, a retrieval-augmented multi-agent system addressing multiple drug discovery tasks. CLADD includes specialized teams for molecular annotation, knowledge graph querying, and prediction synthesis. Evaluations spanned property-specific captioning (BBBP, SIDER, ClinTox, BACE), target identification (DrugBank, KIBA), and toxicity classification. All agents were instantiated with *GPT-4o-mini*, showcasing the utility of general-purpose models when combined with structured RAG mechanisms (10).

Song et al. presented *PharmaSwarm*, a hypothesis-driven agent swarm for therapeutic target and compound identification. The architecture orchestrates three specialized agents (*Terrain2Drug*, *Market2Drug*, *Paper2Drug*) and a central evaluator, all integrated via a shared memory and tool-augmented validation layer. Case studies included idiopathic pulmonary fibrosis and triple-negative breast cancer, combining omics analysis, literature mining, and market signals. Agents were powered by *GPT-4*, *Gemini 2.5*, and *TxGemma* (11).

Yang et al. proposed *DrugMCTS*, a novel multi-agent drug repurposing system that incorporates Monte Carlo Tree Search (MCTS) with structured agent workflows. Using *Qwen2.5-7B-Instruct* for all agents, the system conducts iterative reasoning across molecule retrieval, analysis, filtering, and protein matching. The framework was benchmarked on DrugBank and KIBA, achieving up to 55.34% recall. A case study involving Equol and CXCR3 showed successful prediction of interaction, supported by AutoDock Vina simulations with a binding score of −8.4 kcal/mol (12).

*https://github.com/ChatMED/Prompt-to-Pill

© The Author . Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Inoue et al. introduced *DrugAgent*, an explainable multi-agent reasoning system for drug repurposing. Their architecture coordinates agents handling knowledge graph queries, machine learning scoring, and biomedical literature summarization. Evaluation on a kinase inhibitor dataset revealed strong interpretability and modularity. Detailed ablation studies confirmed that each agent contributes distinctly to the performance. The system employed *GPT-4o*, *o3-mini*, and *GPT-4o-mini*, and the full pipeline is available open-source (13). Among the surveyed systems, only DrugAgent provides a publicly accessible implementation[1].

None of the described multi-agent systems engages with clinical trial simulation, real-world evidence (RWE), or electronic health records (EHRs), thereby limiting their applicability to the preclinical stage of drug development.

This paper presents a systematic analysis of 51 LLM-based studies published between 2022 and 2025, with the goal of proposing a unified multi-agent system, *Prompt-to-Pill*, that spans both preclinical and clinical phases of drug development. The architecture comprises a suite of specialized LLM agents tasked with roles such as molecular design, docking simulation, property prediction, safety filtering, clinical trial emulation, and patient stratification. These agents are coordinated by a dedicated Orchestrator LLM, enabling iterative communication and decision-making across the drug development continuum. Unlike prior frameworks confined to molecule-level reasoning, *Prompt-to-Pill* establishes an integrated pipeline from molecular ideation to virtual trial execution, demonstrating how modular LLM agents can operate synergistically within a closed-loop drug discovery and development ecosystem. A complete implementation of the pipeline is available at GitHub[2].

The following sections explain the systematic approach to choosing the models (Section 2), and how these models are being mapped across the proposed drug discovery pipeline (Section 3). Eventually, the paper presents a discussion where these challenges are contextualized within the broader landscape of biomedical AI (Section 4), and a conclusion (Section 5) where an integrative framework is proposed to advance toward trustworthy, end-to-end LLM-driven drug development.

Methods

The systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The PRISMA framework was employed to ensure transparency, methodological rigor, and reproducibility in identifying, screening, and synthesizing eligible studies. A structured multi-stage review process was followed, encompassing database search, eligibility screening, full-text assessment, and data extraction. The complete selection workflow is detailed in the accompanying PRISMA flow diagram depicted in Figure 1.

Information Sources and Search Strategy

A structured and comprehensive literature search was conducted to identify and evaluate LLM-based approaches applied in drug design and discovery. Four major academic and scientific databases were queried: PubMed, ScienceDirect, Google Scholar, and Springer Nature Link. The search covered the publication period 2022 - 2025.

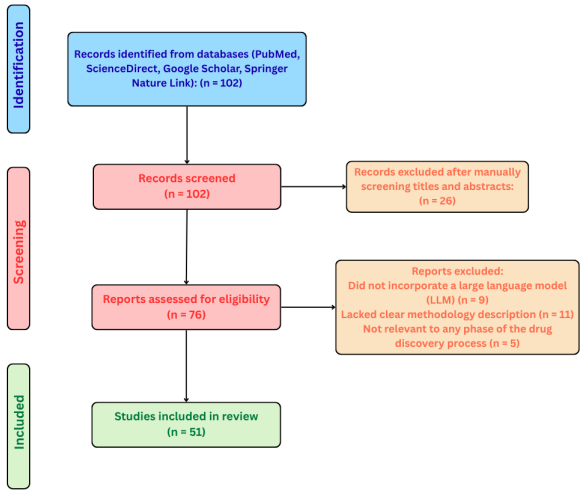

Fig. 1. PRISMA-based selection process.

Search queries were constructed using predefined Boolean combinations, tailored to capture studies addressing different stages of the drug discovery pipeline. Representative search strings included: *"large language models" AND ("target identification" OR "binding site prediction"), "large language models" AND ("molecule generation" OR "de novo molecule generation"), "large language models" AND ("clinical trial design" OR "eligibility criteria extraction" OR "trial outcome prediction"), "retrieval-augmented generation" AND ("drug discovery" OR "clinical trials"), "large language models" AND ("patient recruitment" OR "clinical trial matching"* These terms were selected to align with a conceptual pipeline spanning both preclinical and clinical phases of pharmaceutical development.

Study Selection Process

Two reviewers independently screened the titles and abstracts of all retrieved records. Full-text reviews were then performed to assess eligibility based on the predefined inclusion and exclusion criteria. The inclusion criteria were defined as follows:

- Open-source studies.
- Publications or preprints published between 2022 and 2025.
- Studies written in English.
- Research that incorporates large language models for drug discovery tasks.
- Models with clearly defined input-output structure, functional purpose, and workflow integration potential.
- Studies with application relevance to at least one key stage of the drug discovery or clinical trial process (e.g., molecule generation, ADMET prediction, trial simulation).

And the exclusion criteria, as follows:

- Articles not written in English.
- Studies lacking a clear methodological or architectural description.
- Studies that are not publicly accessible.

[1] https://anonymous.4open.science/r/DrugAgent-B2EA
[2] https://github.com/ChatMED/Prompt-to-Pill

- Research not directly applicable to any stage of drug discovery or development.

The PRISMA flow diagram in Figure 1 details the number of records identified, screened, excluded (with reasons), and finally taken into consideration for building the Prompt-to-Pill pipeline.

Data Extraction and Synthesis

For each included study, detailed metadata were manually extracted into structured tables, one for preclinical models and another for clinical applications. Metadata fields were designed to support both technical evaluation and contextual information from each source as follows:

- **Bibliographic**: Authors, Year, Title, DOI.
- **Technical**: Base model (e.g., GPT-4, BioGPT), Task Type, RAG usage, Evaluation Metrics, Datasets.
- **Reproducibility**: GitHub/Hugging Face links, Input/Output examples.
- **Contextual**: Task Narrative, Clinical Trial Phase (I–IV), Abstract Summary.

Extracted data were then synthesized by stage as needed for the drug development pipeline. Studies were profiled and compared across multiple dimensions including application scope, base architecture, task type, and dataset diversity. This structured comparison informed the construction of the Prompt-to-Pill multi-agent framework introduced later in the paper.

Methodology

Prompt-to-Pill Architecture Foundation

The systematic review of 51 publications (2022–2025) reveals a rapid rise in research activity and diversification of models, tasks, and datasets, peaking in 2024 (14 preclinical, 9 clinical), followed by 16 studies in 2025, and lower counts in 2022 (4 studies) and 2023 (8 studies).

The application of LLMs in drug discovery reveals clear phase-specific trends in architecture, task focus, and data usage. Preclinical studies mainly relied on generative models such as *LLaMA* (14) and *GPT-2* (15) to support tasks that benefit from chemical creativity and open-access datasets like Therapeutics Data Commons (TDC) and DrugBank. These tasks primarily include de novo molecule generation where models like *DrugGen* (2) and *DrugGPT* (3) generate SMILES strings from protein sequences, 3D structures or text-based molecular descriptions (16; 17). Advanced models incorporate spatial constrains (e.g., *3DSMILES-GPT*) or generate RNA therapeutics (e.g., *GenerRNA* (18)), enabling broader design capabilities and iterative compound optimization (19). LLMs also support property estimation and chemical intelligence by predicting ADMET characteristics, synthetic feasibility, and molecular reactivity from either structured input or natural language descriptions (4; 20; 21; 22). Other studies employ LLMs to model biological targets and molecular interactions, enabling hypothesis generation based on biomedical texts and reasoning over drug–gene or protein–protein interactions (23; 24). Finally, some models are used for drug repurposing and combination reasoning, where LLMs identify synergistic therapies or prioritize existing compounds for new indications using patient-specific or knowledge graph data (25; 26).

In contrast to preclinical studies, clinical studies predominantly employed discriminative or hybrid models such as *GPT-4* (27) and *BioBERT* (28), often trained on structured datasets—most notably ClinicalTrials.gov. Among the 20 clinical publications, 10 proposed models that apply across all four clinical trial phases, addressing complex, cross-phase tasks like patient selection, outcome prediction, and document generation. LLMs assist in patient-centered reasoning by aligning unstructured patient records with trial eligibility criteria and providing rationale-based inclusion/exclusion decisions (7; 8). In trial design and simulation, models such as *MediTab* (29) and *CTP* (30) are used to draft protocols, simulate synthetic trials, and forecast outcomes (6). Additionally, LLMs are increasingly deployed for document generation and pharmacovigilance, including tools like *AskFDALabel* (31) and *DAEDRA* (32) that process adverse event reports or regulatory texts, sometimes enhanced by RAG pipelines for context-aware document creation and structured data transformation (33; 34).

Among the reviewed studies, retrieval-augmented generation (RAG) (35) showed limited adoption, despite its potential for complex, multi-modal reasoning tasks. To further assess the reproducibility of existing models, we examined the availability of source code and example input/output data (Table 1). While input examples were provided in most preclinical and clinical studies, output examples were less frequently included. These gaps highlight the need for greater transparency and more detailed explanations of how the models operate.

Based on this empirical analysis we were able to make a strong foundation for the proposed Prompt-to-Pill architecture, a concept that is later implemented by using the AutoGen (36) programming framework for building scalable multi-agent AI systems.

Each agent chosen to be part of the Prompt-to-Pill pipeline is based on or adapted from models that have been rigorously evaluated in their respective domains. Table 2 summarizes the reported performance metrics from the original studies.

Table 1. Availability of I/O Examples, RAG Usage, and Code Links in Studies

Category	Preclinical			Clinical		
	Yes	No	N/A	Yes	No	N/A
GitHub/HF link	19	12	–	10	10	–
Input examples	16	2	1	8	2	–
Output examples	6	12	1	3	7	–
RAG usage	3	28	–	6	14	–

The Prompt-to-Pill Multi-Agent Pipeline

Constructed from the models identified and reviewed in this study, a comprehensive AI-driven pipeline for drug discovery and clinical trial simulation is presented in Figure 2, structured into two main phases: Preclinical Agents and Clinical Agents, both coordinated by a central *Orchestrator*. The workflow is designed to automate the discovery, evaluation, and testing of novel drug candidates - from protein interaction analysis to virtual patient trials - through a collection of specialized agents governed by intelligent coordination logic.

The pipeline begins with a protein target, identified by its UniProt ID, which serves as the input for the *Drug Generation Agent*. To better reflect the end-to-end drug development workflow, the target Dipeptidyl Peptidase 4 (DPP4) was selected as a representative use case, through which the

Table 2. Evaluation metrics of the core agents used in the clinical workflow

Citation	Agent	Dataset / Task	Metric(s)	Value(s)	Type
(2)	Molecule Generation (Druggen)	–	Validity, Novelty, Diversity	99.9%, 41.88%, 60.32%	Generation
(4)	Property Prediction (TxGemma)	BBB Martins	AUROC	0.864	Classification
		ToxCast	AUROC	0.784	Classification
		Lipophilicity	MAE	0.610	Regression
		PPBR	MAE	9.266	Regression
		Solubility	MAE	0.961	Regression
(29)	Trial Outcome Prediction (Meditab)	Phase I Trials	AUROC, PRAUC	0.699, 0.726	Classification
(8)	Patient Matching (Panacea)	SIGIR	BACC, F1, Recall, Precision	0.43, 0.57, 0.52, 0.66	Classification
		TREC 2021	BACC, F1, Recall, Precision	0.47, 0.58, 0.54, 0.69	Classification

Source: Metrics are based on results reported in the original publications.

[1]BACC: Balanced Accuracy; F1: F1 Score; AUROC: Area Under ROC Curve; PRAUC: Area Under Precision-Recall Curve.

[2]MAE: Mean Absolute Error.

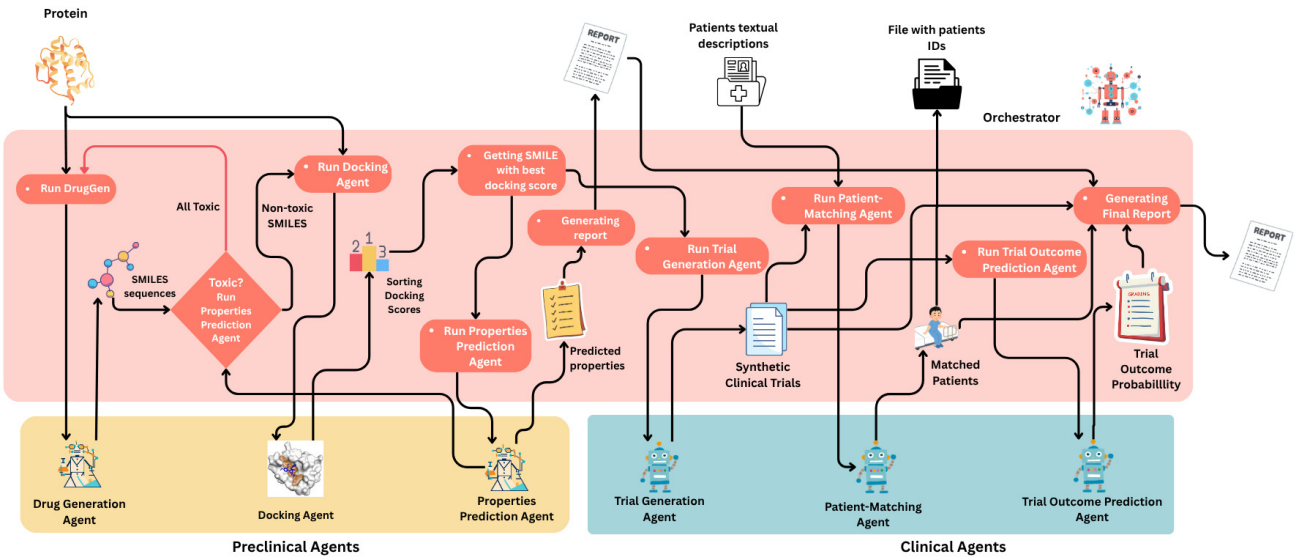

Fig. 2. Prompt-to-Pill Multi-Agent Architecture.

entire pipeline, from molecular design to clinical simulation, is illustrated.

The *Drug Generation Agent* uses a generative molecular language model, based on the DrugGen framework (2), to produce a set of candidate SMILES sequences. These molecules are then passed to the *Properties Prediction Agent*, which evaluates their toxicity profiles using a foundation model adapted for ADMET prediction, inspired by the TxGemma (4). Molecules flagged as toxic are filtered out at this stage.

The non-toxic candidates are subsequently evaluated by the *Docking Agent*, which performs virtual docking against the crystal structure of DPP-4 (PDB ID: *2QT9*), a well-studied dipeptidyl peptidase involved in glucose metabolism. To define the binding region, a grid box is calculated automatically using the coordinates of the ligand co-crystallized in the 2QT9 structure (residue 524 in chain A). To construct the grid, only atoms belonging to the bound ligand are considered. The center of the box is computed by averaging the x, y, and z coordinates of all ligand atoms, effectively placing the grid at the geometric center of the known binding site. The size of the grid is determined by measuring the spatial extent of the ligand in each dimension (i.e., the range between minimum and maximum coordinates in x, y, and z). For this case study, the resulting grid box is cubic, with dimensions of 22.38 Å per side, centered at 41.20x51.10x35.75. The position of the ligand that is chosen by the multi-agent system in the target and the grid box is presented in Figure 3.

Further in the pipeline the molecule with the highest binding affinity is selected, and clinical properties like Blood-Brain Permeability (binary classification: "Crosses BBB" or "Does not cross BBB"), Toxicity (binary classification:"Toxic" or "Non-toxic"), Lipophilicity (normalized score: 000–1000),

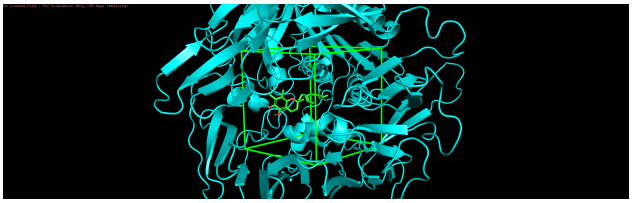

Fig. 3. Visualization of the docking grid (green box) centered on the co-crystallized ligand within the DPP-4 binding pocket (PDB ID: 2QT9, chain A).

Solubility (normalized score: 000–1000), and Plasma Protein Binding Rate (normalized score: 000–1000) are predicted and evaluated with the *Properties Prediction Agent*. Then the whole knowledge is represented as a summary report and the system is stepping in the next stage: the Clinical Phase.

In the Clinical Phase, the *Trial Generation Agent* constructs a Phase I trial protocol tailored to the compound and disease. This protocol is parsed into structured data and passed to the *Patient-Matching Agent*, which employs the *Panacea* model (8) to evaluate patient EHR descriptions and identify candidates who meet the trial's inclusion and exclusion criteria. The agent returns number of matched patients in the final report, and a set of matched patient IDs. These identifiers are saved to a file and the total number of matched patients is computed and forwarded for inclusion in the final trial report. Subsequently, the *Trial Outcome Prediction Agent* uses *MediTab* to assess the likelihood of trial success, taking into account the matched patient cohort, trial design, and the compound's clinical profile.

Finally, the matched patient data, drug properties, and trial design are provided to the *Orchestrator*, which aggregates all outputs into a structured report.

The input and output format for the Dipeptidyl peptidase 4 (DPP4) target are shown in Figure 4.

Discussion

While the rapid integration of LLMs into drug discovery and clinical research has opened transformative opportunities, it has also brought to light a series of critical limitations that must be addressed to fully realize their potential. These systems are increasingly present across the biomedical landscape, yet their adoption raises fundamental questions regarding reasoning, evaluation, and reproducibility.

One of the most pressing challenges is the limited reasoning ability of current models. In fields like pharmacy and medicine, producing the correct output is not sufficient; understanding the cause of a problem is often more important than simply offering a solution. Decisions in these domains must be grounded in clear, justifiable reasoning so that outputs can be critiqued, validated, or overridden by experts. While LLMs excel at generating answers, they often lack the capacity to explain the rationale behind them. To address this, several recent models, such as *AutoTrial*, *TxGemma*, *Y-Mol*, and *MolecularGPT* have introduced mechanisms to make reasoning more explicit. These include retrieval-augmented generation (4; 6; 37), instruction-tuned multitask learning (5; 38), and multi-hop rationale generation (37; 39). Such approaches represent important progress toward interpretability. However, reasoning in LLMs remains inherently probabilistic and highly prompt-dependent. Without standardized methods to evaluate the quality or consistency of this reasoning, especially in high-stakes biomedical applications, trust in model outputs remains limited.

Another major challenge in LLM-based drug discovery is the inconsistency in evaluation protocols across model types. Generative models are assessed with varied metrics like validity, docking scores, or QED (2; 16; 18; 40), but lack a unified benchmark. Discriminative models, used for tasks such as toxicity or ADMET prediction, report metrics such as AUROC or F1 scores (4; 5; 38), yet differ in datasets and thresholds. Knowledge-retrieval and reasoning systems often rely on qualitative outputs without standardized measures (37; 39). This fragmentation hinders comparability and progress. To address this, the field urgently needs task-specific, model-type-sensitive benchmarks. Generative models should be evaluated on fixed panels of docking targets and synthesis metrics, discriminative models on shared datasets and criteria, and retrieval models on precision and evidence alignment.

Beyond evaluation, reproducibility remains an unresolved concern. Many of the studies surveyed in this review do not provide access to source code, pre-trained models, or even illustrative input-output pairs. Even when repositories are available, documentation is often sparse or inconsistent, limiting the ability of others to replicate results or integrate models into multi-agent systems. This gap stifles innovation, wastes resources, and obscures scientific transparency.

These models are the future of drug development, but there is still much work to be done. In short, the path forward requires not only better models, but better systems around the models. This includes standardized evaluations, transparent documentation, expert-guided development, and thoughtful regulation. Only by meeting these unmet needs can we ensure that LLMs evolve from experimental tools to trusted agents in the future of biomedical discovery.

Conclusion

Through a structured review of 51 recent studies (2022–2025), the application of LLMs across all phases of the drug design and discovery process was examined from early-stage molecular generation and property estimation to patient–trial matching and clinical outcome forecasting.

Important trends were revealed through a cross-phase analysis. In preclinical studies, generative models were favored for creative tasks such as de novo molecule design, often utilizing open-access chemical datasets and scoring functions. In contrast, clinical applications were found to rely more heavily on discriminative and hybrid models for structured data interpretation and safety evaluation, with a growing interest observed in RAG-based systems to support explainability and real-world deployment. Despite these advancements, the field was noted to remain fragmented in terms of evaluation standards, reproducibility, and reasoning transparency. Many models were reported to lack shared benchmarks or accessible code and examples, factors that hinder both adoption and trust.

To illustrate the current capabilities of LLMs, a multi-agent system architecture was proposed. The "Prompt-to-Pill" framework was designed to integrate specialized agents and demonstrate how coordinated LLM workflows can be employed to automate complex decisions across preclinical and clinical stages. The system reflects not only the functional diversity of

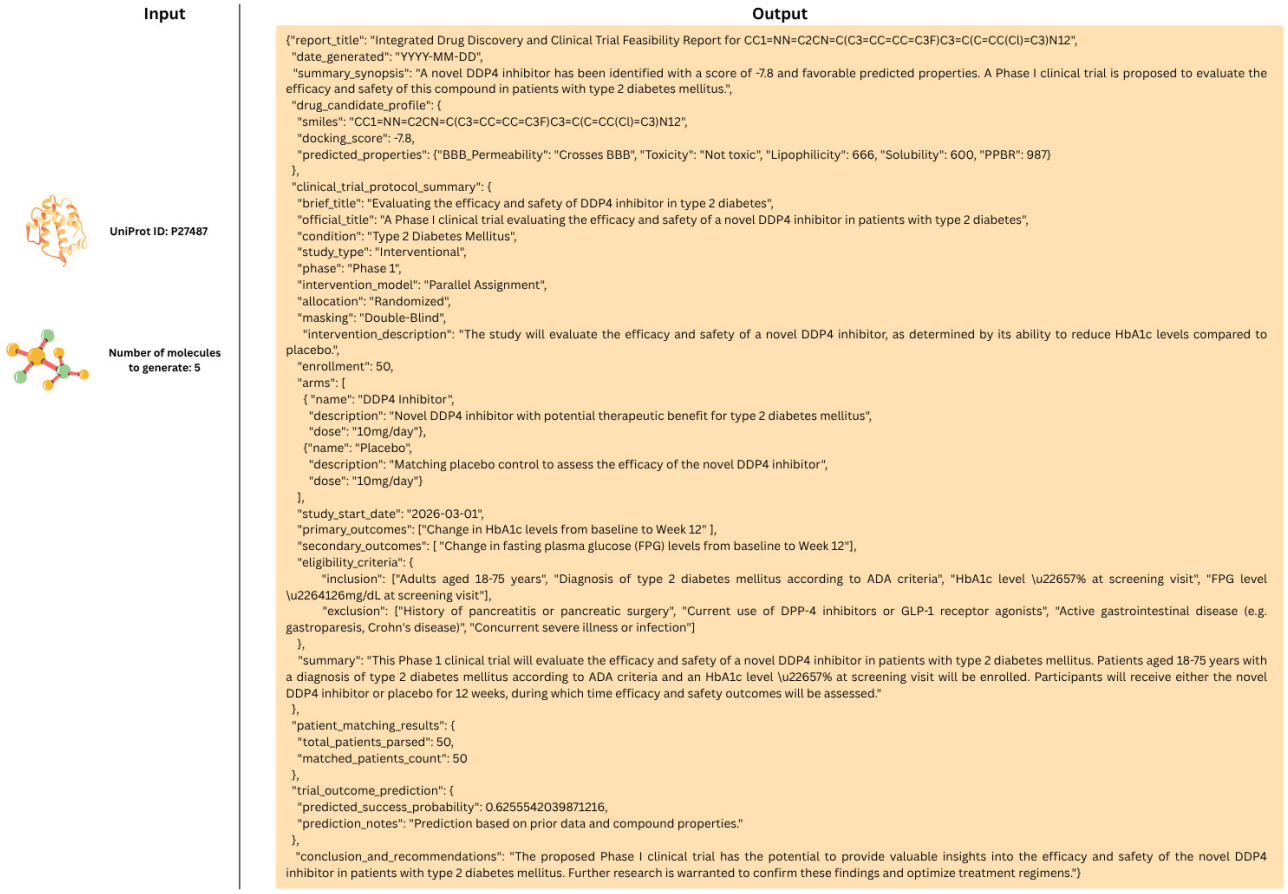

Fig. 4. Prompt-to-Pill's I/O example.

modern LLMs but also their capacity to collaborate, iterate, and self-correct within an intelligent, modular design.

Looking ahead, it is anticipated that the future of LLM-based drug discovery will depend not only on the development of more capable models but also on the systems and standards that support their use. Progress will require the establishment of improved evaluation protocols, causal reasoning frameworks, open-source transparency, and appropriate regulatory oversight. If these challenges are addressed, LLMs may not only accelerate the pace of drug discovery, they may fundamentally redefine it.

Code Availability

The full implementation of the Prompt-to-Pill multi-agent drug discovery and clinical simulation pipeline is available at the GitHub repository: `https://github.com/ChatMED/Prompt-to-Pill`.

Funding

Funded by the European Union under Horizon Europe project (ChatMED Grant Agreement ID: 101159214). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

1. David Oniani, Jordan Hilsman, Chengxi Zang, Junmei Wang, Lianjin Cai, Jan Zawala and Yanshan Wang Emerging Opportunities of Using Large Language Models for Translation Between Drug Molecules and Indications *arXiv preprint arXiv:2402.09588*, 2024.
2. Mahsa Sheikholeslami, Navid Mazrouei, Yousof Gheisari, Afshin Fasihi, Matin Irajpour and Ali Motahharynia. DrugGen enhances drug discovery with large language models and reinforcement learning. *Scientific Reports*, 15, 13445, 2025.
3. Yuesen Li, Chengyi Gao, Xin Song, Xiangyu Wang, Yungang Xu and Suxia Han. DrugGPT: A GPT-based Strategy for Designing Potential Ligands Targeting Specific Proteins. *bioRxiv (Cold Spring Harbor Laboratory)*, 2023.
4. Eric Wang, Samuel Schmidgall, Paul F. Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet and Shekoofeh Azizi. TxGemma: Efficient and Agentic LLMs for Therapeutics. *arXiv preprint arXiv:2504.06196*, 2025.
5. Yuyan Liu, Sirui Ding, Sheng Zhou, Wenqi Fan and Qiaoyu Tan. MolecularGPT: Open Large Language Model (LLM) for Few-Shot Molecular Property Prediction. *arXiv preprint arXiv:2406.12950*, 2024.

6. Zerui Xu, Fang Wu, Yuanyuan Zhang and Yue Zhao. Retrieval-Reasoning Large Language Model-based Synthetic Clinical Trial Generation. *arXiv preprint arXiv:2410.12476*, 2025.
7. Surabhi Datta, Kyeryoung Lee, Liang-Chin Huang, Hunki Paek, Roger Gildersleeve, Jonathan Gold, Deepak Pillai, Jingqi Wang, Mitchell K. Higashi, Lizheng Shi, Percio S. Gulko, Hua Xu, Chunhua Weng and Xiaoyan Wang. Patient2Trial: From patient to participant in clinical trials using large language models. *Informatics in Medicine Unlocked*, 101615, 2025.
8. Jiacheng Lin, Hanwen Xu, Zifeng Wang, Sheng Wang and Jimeng Sun. Panacea: A foundation model for clinical trial search, summarization, design, and recruitment. *arXiv preprint arXiv:2407.11007*, 2024.
9. Gao, B., Bai, H., and Zhang, P. [Proposal] A Multi-Agent Framework for Reliable Drug-Target Interaction Prediction. Tsinghua University Course: Advanced Machine Learning.
10. Lee, N., De Brouwer, E., Hajiramezanali, E., Biancalani, T., Park, C., and Scalia, G. (2025). RAG-Enhanced Collaborative LLM Agents for Drug Discovery. *arXiv preprint*, arXiv:2502.17506.
11. Song, K., Trotter, A., and Chen, J. Y. (2025). LLM agent swarm for hypothesis-driven drug discovery. *arXiv preprint*, arXiv:2504.17967.
12. Yang, Z., Wan, Y., Li, Y., Matsuda, Y., Xie, T., and Song, L. (2025). DrugMCTS: A drug repurposing framework combining multi-agent, RAG and Monte Carlo Tree Search. *arXiv preprint*, arXiv:2507.07426.
13. Inoue, Y., Song, T., and Fu, T. (2024). DrugAgent: Explainable drug repurposing agent with large language model-based reasoning. *arXiv e-prints*, arXiv:2408.13378.
14. Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave and Guillaume Lample. LLAMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023.
15. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9, 2019.
16. Jike Wang, Hao Luo, Rui Qin, Mingyang Wang, Xiaozhe Wan, Meijing Fang, Odin Zhang, Qiaolin Gou, Qun Su, Chao Shen, Ziyi You, Liwei Liu, Chang-Yu Hsieh, Tingjun Hou and Yu Kang. 3DSMILES-GPT: 3D molecular pocket-based generation with token-only large language model. *Chemical Science*, 16:637-648 , 2024.
17. Carl Edwards, Tuan Lai1,, Kevin Ros, Garrett Honke, Kyunghyun Cho and Heng Ji. Translation between Molecules and Natural Language. *arXiv preprint arXiv:2204.11817*, 2022.
18. Yichong Zhao, Kenta Oono, Hiroki Takizawa and Masaaki Kotera. GenerRNA: A generative pre-trained language model for de novo RNA design. *PLoS ONE*, 19(10): e0310814, 2024.
19. Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu and Xiangxiang Zeng. DrugAssist: a large language model for molecule optimization. *Briefings in Bioinformatics*, 26, 2024.
20. Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, Kai Yu and Xin Chen. Developing ChemDFM as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6:102523, 2025.
21. Juan Manuel Zambrano Chaves, Eric Wang, Tao Tu, Eeshit Dhaval Vaishnav, Byron Lee, S. Sara Mahdavi, Christopher Semturs, David Fleet, Vivek Natarajan and Shekoofeh Azizi. Tx-LLM: A Large Language Model for Therapeutics. *arXiv preprint arXiv:2406.06316*, 2024
22. He Cao, Zijing Liu, Xingyu Lu, Yuan Yao and Yu Li. InstructMol: Multi-Modal Integration for Building a Versatile and Reliable Molecular Assistant in Drug Discovery. *arXiv preprint arXiv:2311.16208*, 2024.
23. Jon-Michael T. Beasley, Kara Schatz, Elvin Ding, Marcello DeLuca, Nahed Abu Zaid, Nyssa N. Tucker, Rada Y. Chirkova, Daniel J. Crona, Alexander Tropsha and Eugene N. Muratov. TARRAGON: Therapeutic Target Applicability Ranking and Retrieval-Augmented Generation Over Networks. /em bioRxiv (Cold Spring Harbor Laboratory), 2025.
24. Ziwen Li, Xiang 'Anthony' Chen and Youngseung Jeon. GraPPI: A Retrieve-Divide-Solve GraphRAG Framework for Large-scale Protein-protein Interaction Exploration. *arXiv preprint arXiv:2501.16382*, 2025.
25. Carl Edwards, Aakanksha Naik, Tushar Khot, Martin Burke, Heng Ji, Tom Hope. SynerGPT: In-Context Learning for Personalized Drug Synergy Prediction and Drug Design. *bioRxiv (Cold Spring Harbor Laboratory)*, 2023.
26. Rico Andre Schmitt, Konstantin Buelau, Leon Martin, Christoph Huettl, Michael Schirner, Leon Stefanovski and Petra Ritter Biological Database Mining for LLM-Driven Alzheimer's Disease Drug Repurposing. *bioRxiv (Cold Spring Harbor Laboratory)*, 2024.
27. OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2024.
28. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234–40, 2019.
29. Zifeng Wang, Chufan Gao, Cao Xiao and Jimeng Sun. MediTab: Scaling Medical Tabular Data Predictors via Data Consolidation, Enrichment, and Refinement. *arXiv preprint arXiv:2305.12081*, 2024.
30. Michael Reinisch, Jianfeng He, Chenxi Liao, Sauleh Ahmad Siddiqui and Bei Xiao. CTP-LLM: Clinical Trial Phase Transition Prediction Using Large Language Models. *arXiv preprint arXiv:2408.10995*, 2024.
31. Leihong Wu, Hong Fang, Yanyan Qu, Joshua Xu and Weida Tong. Leveraging FDA Labeling Documents and Large Language Model to Enhance Annotation, Profiling, and Classification of Drug Adverse Events with AskFDALabel. *Drug Safety*, 2025.
32. Chris von Csefalvay. DAEDRA: A language model for predicting outcomes in passive pharmacovigilance reporting. *arXiv preprint arXiv:2402.10951*, 2024.
33. Nigel Markey, Ilyass El-Mansouri, Gaetan Rensonnet, Casper van Langen, and Christoph Meier. From RAGs to riches: Utilizing large language models to write documents for clinical trials. *Clinical Trials*, 2025.
34. Jeffery L Painter, Venkateswara Rao Chalamalasetti, Raymond Kassekert and Andrew Bate. Automating pharmacovigilance evidence generation: using large language models to produce context-aware structured query language *JAMIA Open* 8, 2024.

35. Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*, 2024.
36. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al. (2024). Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *Proceedings of the First Conference on Language Modeling*.
37. Yichun Feng, Jiawei Wang, Ruikun He, Lu Zhou and Yixue Li. A Retrieval-Augmented Knowledge Mining Method with Deep Thinking LLMs for Biomedical Research and Clinical Support. *arXiv preprint arXiv:2503.23029*, 2025.
38. Tengfei Ma, Xuan Lin, Tianle Li, Chaoyi Li, Long Chen, Peng Zhou, Xibao Cai, Xinyu Yang, Daojian Zeng, Dongsheng Cao and Xiangxiang Zeng. Y-Mol: A Multiscale Biomedical Knowledge-Guided Large Language Model for Drug Development. *arXiv preprint arXiv:2410.11550*, 2024.
39. Zifeng Wang, Cao Xiao and Jimeng Sun AutoTrial: Prompting Language Models for Clinical Trial Design *arXiv preprint arXiv:2305.11366*, 2023.
40. Jie Yue, Bingxin Peng, Yu Chen, Jieyu Jin, Xinda Zhao, Chao Shen, Xiangyang Ji, Chang-Yu Hsieh, Jianfei Song, Tingjun Hou, Yafeng Deng and Jike Wang. Unlocking comprehensive molecular design across all scenarios with large language model and unordered chemical language. *Chemical Science*, 15:13727–40, 2024.

Ivana Vichentijevikj Miss Vichentijevikj is a postgraduate student in Bioinformatics with a biomedical background. Her academic and research interests focus on the application of computational methods in drug discovery, drug development, and personalized medicine. She is particularly interested in leveraging gene expression data, molecular modeling, and machine learning to support therapeutic development and target identification. Ivana is a part of the ChatMED project, where she contributes to the integration of generative AI tools into clinical and pharmaceutical workflows. Her work reflects a strong dedication to advancing translational bioinformatics and promoting the role of AI in the future of healthcare systems.

Kostadin Mishev. Prof. Mishev is a professor of Computer Science and Engineering and an expert in the emerging interdisciplinary field that bridges artificial intelligence and DevOps, AIOps, bringing cutting-edge methodologies for operationalizing machine learning pipelines in real-world environments. As a member of the Scientific Committee of the ChatMED project, he contributes strategic and technical expertise in deploying robust, compliant, and scalable AI-driven systems within clinical workflows. Prof. Mishev is also a key industry representative. His portfolio includes the development and deployment of production-grade AI applications, notably, he has led the design of VoiceBot technologies tailored for medical contexts, production-ready for integrating into clinical decision support systems.

Monika Simjanoska Misheva. Dr. Simjanoska Misheva is a professor of Computer Science and Engineering with a research focus at the intersection of artificial intelligence, bioinformatics, and clinical decision support systems. She is currently the coordinator of the ChatMED project, a pioneering initiative that aims to integrate generative AI across all levels of a national healthcare system. Her work is driven by the ambition to ensure that the deployment of AI in healthcare is not only innovative but also compliant with the latest European regulations. Dr. Simjanoska's contributions extend across multi-agent systems, large language models, and their applications in neurology, cancer diagnosis, and personalized medicine. Her recent efforts center on creating regulatory-aware AI architectures capable of multimodal reasoning with genomic, imaging, and biosignal data.