DrugAgent: Multi-agent large language model-based reasoning for drug-target interaction prediction

Yoshitaka Inoue[1,2,3], Tianci Song[1], Xinling Wang[4], Tianfan Fu[5], Augustin Luna[2,3]

1 University of Minnesota, US, 2 National Library of Medicine, US, 3 National Cancer Institute, US

4 Northeastern University, US, 5 Nanjing University, China

Introduction & Conclusion (TLDR)

DrugAgent is a multi-agent system that improves **drug-target interaction (DTI) prediction** combines the strengths of large language models (LLMs), knowledge graphs, and literature search. It uses:

- A **coordinator** to manage five specialized agents.
- **AI, KG, and Search agents** to collect evidence from ML models, biomedical graphs, and web data.
- **Reasoning agent** with CoT & ReAct to integrate and explain results.

DrugAgent achieves **+45% better F1 score** than baseline GPT-4o mini and provides **human-readable reasoning** for each prediction.

- **AI Agent** is the most impactful, while **KG** and **Search** enhance precision and reduce false positives.
- Designed for drug discovery, but extendable to other fields requiring multi-source evidence and explainable AI.

Method

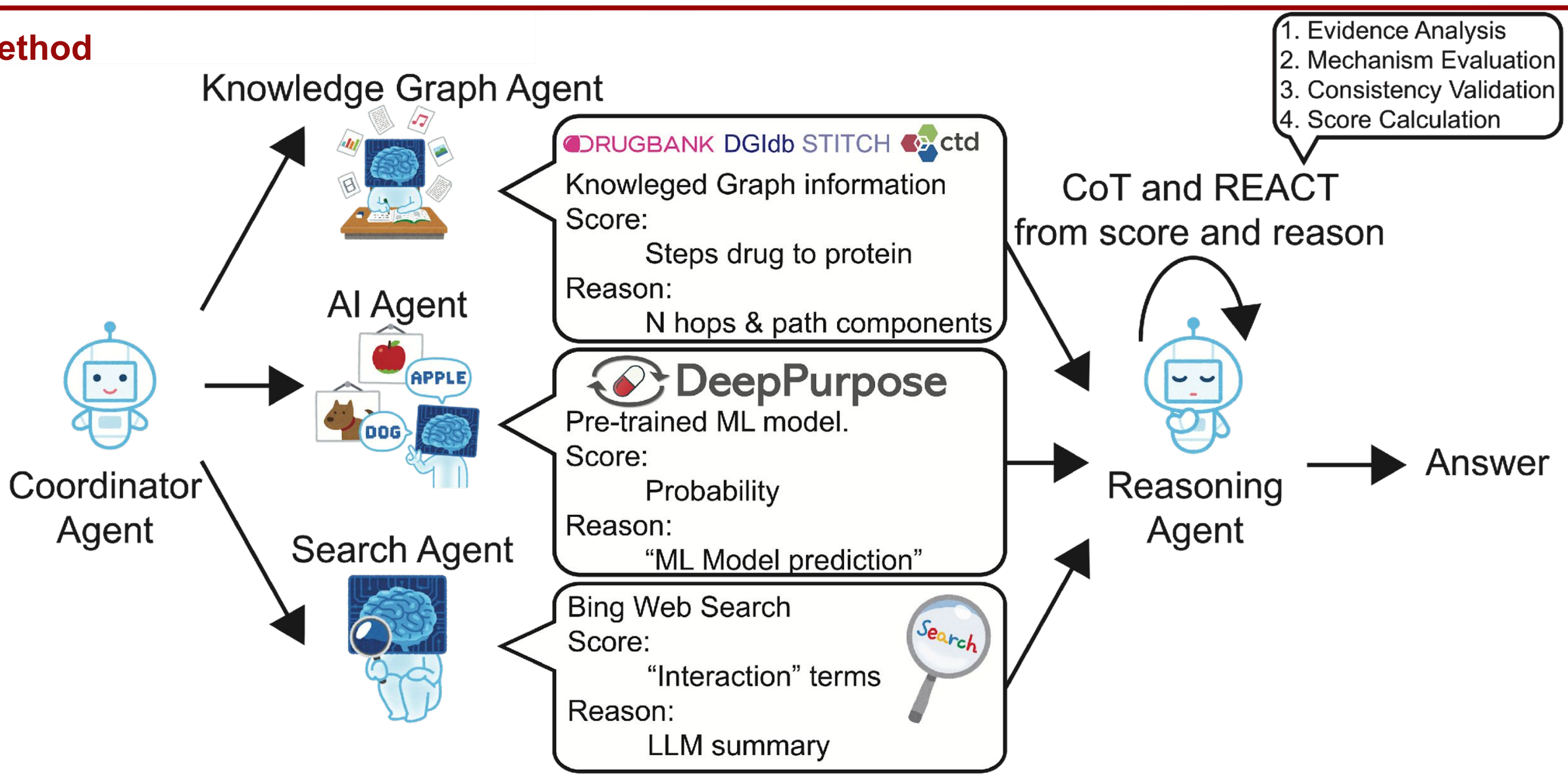

Fig. 1: Multi-agent system architecture for DTI analysis. The system consists of a "Coordinator Agent" that manages three specialized agents for evidence gathering: (1) a Knowledge Graph Agent" accessing biomedical databases (DrugBank, DGIdb, STITCH, CTD) to analyze path-based relationships, (2) an "AI Agent" utilizing the pre-trained DeepPurpose ML model for probability prediction, and (3) a "Search Agent" performing Bing Web Search for literature evidence. The "Reasoning Agent" then integrates this information through CoT and ReAct frameworks to generate final scores with reasoning.

Result

Metric	DrugAgent (o3-mini)	GPT-4o mini	w/o Search	w/o AI	w/o KG
Reasoning	✔	✘	✘	✘	✘
F1 (↑)	0.514 (±0.084)	0.355* (±0.039)	0.481 (±0.037)	0.274* (±0.050)	0.298* (±0.033)
Precision (↑)	0.571 (±0.109)	0.231* (±0.024)	0.338* (±0.028)	0.202* (±0.040)	0.187* (±0.018)
Recall (↑)	0.476 (±0.076)	1.000 (±0.000)	0.982 (±0.002)	0.512 (±0.089)	1.000 (±0.000)
Specificity (↑)	0.978 (±0.000)	0.702* (±0.003)	0.836* (±0.003)	0.765* (±0.003)	0.597* (±0.004)
AUROC (↑)	0.941 (±0.003)	0.938 (±0.002)	0.966 (±0.002)	0.670* (±0.109)	0.953 (±0.003)
AUPRC (↑)	0.677 (±0.102)	0.554 (±0.076)	0.745 (±0.035)	0.456* (±0.082)	0.706 (±0.106)
Runtime (↓)	≈30.000s	≈25.000s	-	-	-
# Tokens (↓)	≈2000–3000	≈2000–3000	-	-	-
Token Cost (↓)	$0.025–$0.037	$0.0015–$0.003	-	-	-

Table 1: Comparison of evaluation metrics across models. Results show means and standard deviations (in brackets) over five independent runs, each sampling 50 subsets. Arrows (↑/↓) indicate better direction, bold indicates best performance, underline describes second best, and * denotes statistical significance (p-value<0.05) compared to DrugAgent.

This research was supported in part by the Division of Intramural Research (DIR) of the National Library of Medicine (NLM), National Institutes of Health (NIH) (ZIALM240126).

ArXiv GitHub

Use Case

Discovering new drug-target interactions with DrugAgent

Input: Drug name (e.g., Imatinib), target protein (e.g., ABL1)
Output:

1. Interaction score: 0.82 (Score range: 0-1, ≥0.5: Likely interaction)
2. Supporting evidence (Reasoning Agent assigns weights)
   - Direct binding confirmed in DrugBank database (KG score: 0.9)
   - Multiple studies support interaction (Search score: 0.8)
   - AI predicts strong binding based on structure (AI score: 0.7)
   - **0.4**×0.9 (KG) + **0.3**×0.8 (Search) + **0.3**×0.7 (AI) = 0.82 (weights in **bold**)

Dataset

We used a kinase-compound activity dataset (Anastassiadis, DOI:10.1038/nbt.2017) for evaluation. The dataset measures kinase activity as the percentage of remaining enzymatic function after compound exposure. The activity range is from 0 to 100. We set the 50 as the threshold to binarize activity values and denote the presence of a drug-target interaction. We randomly selected 50 unique drug-target combinations fed into DrugAgent (using only drug and protein name) and other models 5 times for evaluation.