DrugAgent: Reliable Multi-Agent Aggregation under Conflicting Biomedical Evidence

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

Abstract

Integrating heterogeneous evidence into reliable decisions remains challenging: approaches typically rely on single-modality data or loosely coupled multi-agent prompting and thus lack ways for reconciling conflicts. We present DrugAgent, a multi-agent aggregation framework that separates evidence reasoning from rule-guided final decision-making while integrating outputs from machine learning (ML), knowledge graph (KG), and retrieval-augmented generation (RAG) modules. This design enables interpretable, conflict-aware aggregation. As a case study, we apply DrugAgent to predict drug-target interactions. On a kinase benchmark of 900 pairs spanning 178 kinases and 42 inhibitors, DrugAgent produces outputs judged faithful to the provided evidence (98.8%). Plausibility scores were broadly high across all ground-truth classes. In the DrugAgent setting, 79% of Weak interacting cases, 81% of Moderate cases, and 77% of Strong cases received plausibility scores of 3-4 for the agent returned label explanations; Strong cases were somewhat more likely to receive a score of 5 (15% vs 1% for Weak and 3% for Moderate). This work establishes a benchmark for evaluating reasoning over diverse sources typical in biomedicine, as well as agent capabilities to produce rationales that are human-interpretable and faithful to the source evidence. Code: https://github.com/sciluna/DrugAgent.

[1]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA [2]Computational Biology Branch, National Library of Medicine, Bethesda, MD, USA [3]Khoury College of Computer Sciences, Northeastern University, Arlington, VA, USA [4]Developmental Therapeutics Branch, National Cancer Institute, Bethesda, MD, USA [5]State Key Laboratory for Novel Software Technology at Nanjing University, School of Computer Science, Nanjing University, Nanjing, Jiangsu, China. Correspondence to: Tianfan Fu <futianfan@gmail.com>, Augustin Luna <augustin@nih.gov>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.

1. Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in solving a wide range of problems using human-friendly inputs (Wei et al., 2022a). However, they remain unreliable for scientific tasks that require integrating specialized domain knowledge and multiple viewpoints. To address these limitations, multi-agent systems (Du et al., 2023) have emerged as a practical paradigm in which specialized agents interact and incorporate external tools such as knowledge graphs (KG) (Shu et al., 2024) and retrieval-augmented generation (RAG) (Lewis et al., 2020). In this paper, we present a framework for explainable aggregation of heterogeneous evidence through multi-agent reasoning.

We propose a multi-agent architecture in which each agent specializes in a distinct evidence source for DTI prediction, in which specialized agents collect complementary evidence from ML, KG, and literature. Unlike prior multi-agent pipelines that rely on loosely coupled prompting (Du et al., 2023; Tang et al., 2023) or implicit consensus (Ki et al., 2025; Pitre et al., 2025; Yao et al., 2025), our framework converts agent outputs into a structured intermediate representation and reconciles conflicts through explicit aggregation rules. A Coordinator Agent manages execution and standardized communication across agents, while each specialist agent produces both a categorical label and a concise rationale, enabling the final decision to remain interpretable. Although demonstrated in drug discovery, the framework is broadly applicable to scientific settings that require consistent integration of heterogeneous evidence.

2. Datasets

For evaluation, we used a kinase–compound activity dataset from a large-scale profiling study (Anastassiadis et al., 2011). The original assay reports the percent remaining kinase activity after compound exposure. To formulate an interaction-strength prediction, we discretized this continuous readout into three classes based on inhibition strength: Strong (remaining activity $\leq 25\%$), Moderate (25–50%), and Weak (50–80%), while pairs with $> 80\%$ remaining activity were treated as effectively inactive and excluded. These cutoffs are consistent with the activity ranges used to

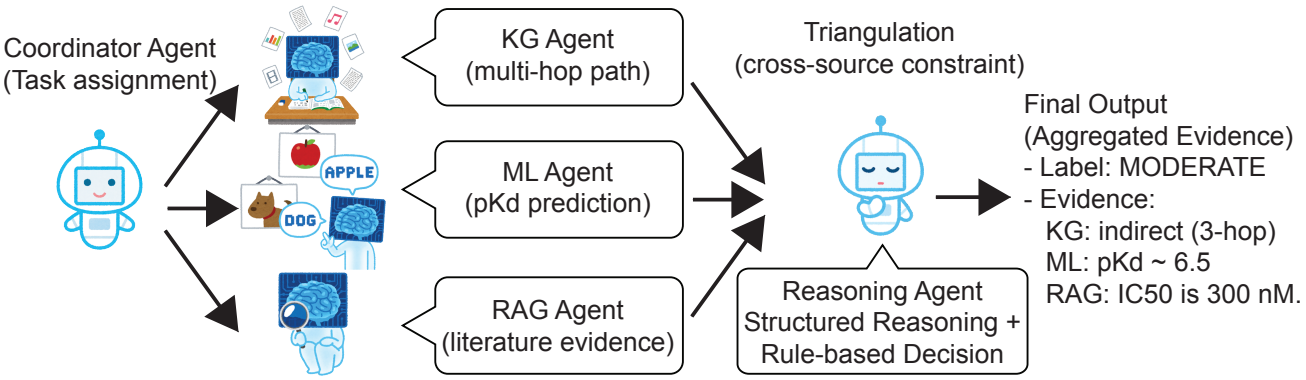

Figure 1. Overview of DrugAgent.

summarize kinase inhibition patterns in the original profiling study (Anastassiadis et al., 2011). From 1,895 labeled interactions, we constructed a balanced benchmark of 900 pairs (300 per class), covering 178 kinases and 42 inhibitors.

3. Methods

3.1. Overview of DrugAgent

3.1.1. Coordination Layer

The coordination layer supports two execution modes. In the default fast mode, ML, KG, and RAG modules are invoked directly as independent components without iterative inter-agent communication. Optionally, a Coordinator Agent can be enabled to decompose the task and orchestrate the execution order across agents. In both modes, module outputs are constrained to a shared JSON schema, ensuring consistent structured inputs to the downstream reasoning stage.

3.1.2. ML Agent

The ML Agent predicts drug-target interactions using DeepPurpose (Huang et al., 2020), a pre-trained model on BindingDB (Liu et al., 2025) that estimates continuous pK_d values from molecular SMILES strings and protein sequences.

3.1.3. KG Agent

We construct a heterogeneous drug-gene KG by integrating curated biomedical resources, including DrugBank (Knox et al., 2024), CTD (Davis et al., 2023), and BindingDB (Liu et al., 2025). We perform multi-hop reasoning over this graph using breadth-first search up to a maximum depth of $H = 5$. Candidate paths are ranked by a structural scoring function that favors shorter paths while penalizing high-degree hub nodes, $S(P) = \frac{1}{|P|} \cdot \lambda^{H(P)}$, where $|P|$ is the path length and $H(P)$ denotes the number of hub nodes (degree > 300) in path P. We set $\lambda = 0.6$.

3.1.4. PubMed RAG Agent

We built an RAG agent over PubMed Central (PMC) Open Access articles. To address sparse direct co-mentions, we use a retrieval strategy with 3 queries: drug-target pairs, drug-only, and target-only. For each query type, we retrieve up to 10 documents from the PMC search application programming interface (API). Query embeddings are generated using the `text-embedding-3-large` model.

3.1.5. Reasoning Agent

The Reasoning Agent operates in two stages: structured reasoning and rule-guided decision-making. In the first stage, outputs from the ML, KG, and RAG modules are converted into a structured JSON format that allows communication reasoning logic between sub-agents. Each source contributes an *evidence_analysis* entry that includes evidence source, agent thoughts, proposed action, and agent observations about the input, along with its proposed DTI label. A summarization step then adds a *summary_reasoning* field that captures agreement, conflict, and uncertainty across sources. This stage summarizes evidence but does not produce the final decision. In the second stage, a decision module produces the final output, including a label and supporting rationale. Decisions follow a rule priority order: majority agreement is applied first, followed by tie-breaking. When these rules are inconclusive, a constrained LLM-based fallback is applied.

4. Results

Faithfulness. We evaluate faithfulness using an LLM-as-a-judge (LLMaJ) with a rubric that scores on a 1-5 scale based on whether the explanation (i) is grounded in the input, (ii) contains contradictions from input to output, (iii) introduces unsupported claims, or (iv) omits critical information. The LLMaJ prompt requests Boolean indicators for these four criteria. Faithfulness scores were high across all configurations, with most samples rated as *Agree* or *Strongly Agree*,

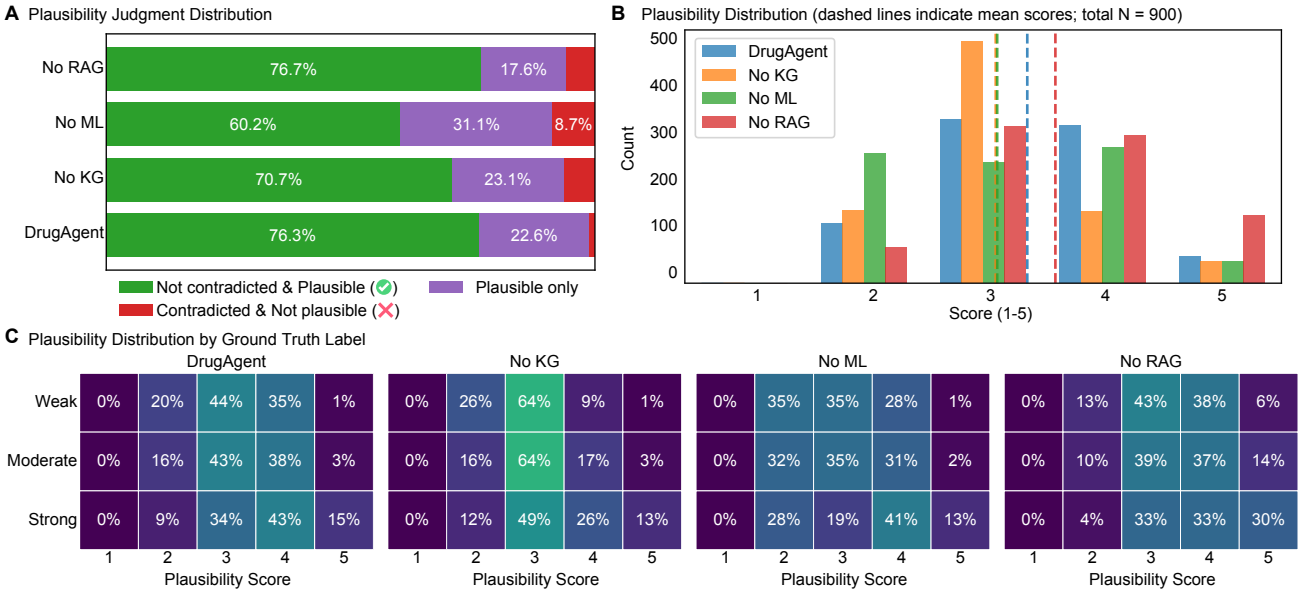

Figure 2. Plausibility-based evaluation across model variants. (A) Distribution of LLM-as-a-judge plausibility outcomes for DrugAgent and ablated variants, showing the proportions of Not contradicted & Plausible, Plausible only, and Contradicted & Not plausible. (B) Distribution of plausibility scores (1–5) across variants; dashed vertical lines indicate the mean score for each variant. (C) Plausibility score distributions stratified by ground-truth interaction strength (Weak, Moderate, Strong). Each heatmap shows the proportion of samples assigned to each plausibility score within a ground-truth class.

indicating close alignment between the generated explanations and the provided evidence (N for Agree and Strongly Agree; DrugAgent: $N = 889$, No ML: $N = 896$, No KG: $N = 857$, No RAG: $N = 900$; total $N = 900$).

Biological Plausibility. We evaluate biological plausibility using an LLMaJ with a rubric that scores each DTI on a 1-5 scale based on (i) absence of contradiction and (ii) mechanistic coherence.

DrugAgent produces plausibility judgments where 98.9% of outputs (Figure 2A) are either (i) both evidence-grounded and coherent or (ii) plausible only (reasonable but weakly supported). Ablation results suggest that each component contributes to plausibility: removing KG, ML, or RAG increases the proportion of outputs rated as having contradictory evidence and not plausible relative to the DrugAgent model, although the magnitude differs across components.

Removing RAG shifts the plausibility distribution toward higher scores, with the mean increasing relative to DrugAgent (Figure 2B). Counts for scores 4-5 increase, despite the lack of grounding in retrieved evidence. This indicates that higher plausibility can result from less evidence checked and not necessarily improved alignment to the evidence.

For all ablation configurations, scores concentrate in the 3-4 range (Figure 2C), with relatively few low scores (1-2). Increasing evidence generally shifts values toward higher scores, but this shift is similar across Weak, Moderate, and Strong ground-truth labels. Notably, even for Weak interactions, a substantial proportion of samples receive scores of 3-4, and in some cases 5, indicating that higher plausibility does not necessarily distinguish true interaction strength.

This reveals a mismatch between perceived plausibility and evidence-grounded correctness. While plausibility scores are high and increase with reduced constraints, they do not reliably reflect alignment with biological signals. These findings indicate that LLMaJ plausibility captures perceived coherence rather than integration of evidence, highlighting the need to account for a base amount of evidence.

Evidence Combination Consistency and Conflict Dynamics. We next analyze how evidence patterns map to final output labels under the aggregation policy (Figure 3). For each pattern, we compare ground-truth labels with final outputs and track how predictions change through the aggregation pipeline (Figure 3A). Each case is represented as a triplet of module outputs ordered as ML|KG|RAG (e.g., M|W|S). Output labels are nearly deterministic for given evidence patterns, consistent with rule-based decision behavior. Outcomes vary with the degree of agreement: consensus patterns yield confident predictions, whereas full conflict shifts predictions toward Moderate, reflecting uncertainty-aware reconciliation (Pitre et al., 2025) (Figure 3B). Biases appear under conflicting evidence at the pattern level, while the overall bias remains near zero; average 0.02(Figure 3C). Case studies illustrate this mechanism: RAG-derived evidence can dominate final decisions when it offers direct experimental support, but it can also introduce errors if

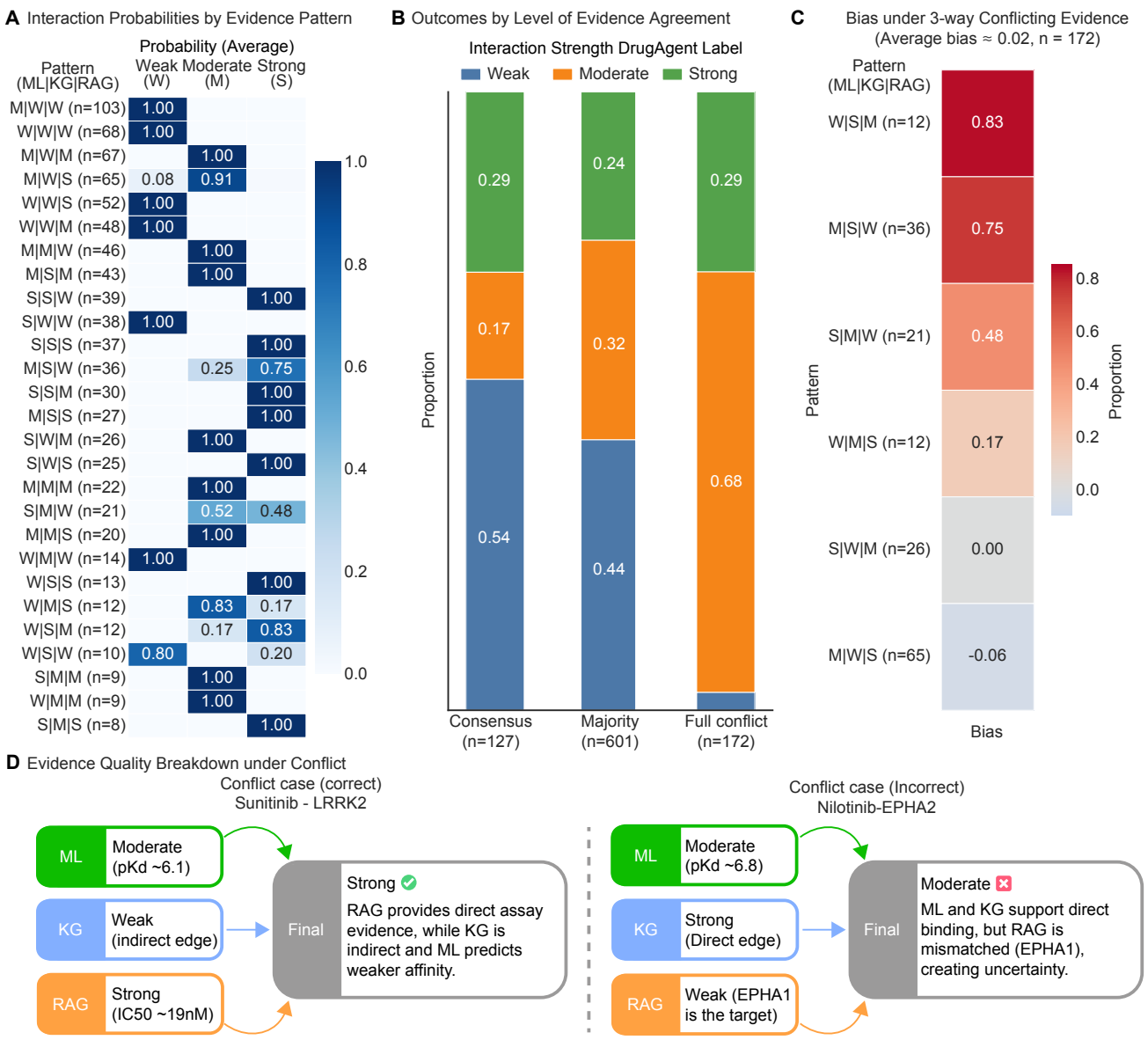

Figure 3. Label-level decision behavior under multi-source evidence aggregation. (A) Pattern-to-decision mapping shows near-deterministic associations between evidence configurations (ML|KG|RAG) and predicted labels. (B) Decision distributions vary with evidence agreement: consensus cases favor confident predictions, whereas full conflict leads to a strong shift toward Moderate outcomes. (C) Under full conflict, directional bias emerges at the pattern level (bias = $P(\text{Strong}) - P(\text{Low})$), although the overall bias remains near zero, indicating globally balanced but locally structured behavior. (D) Representative case studies illustrate the decision mechanism: (left) correct prediction under conflict where direct assay evidence from RAG dominates weaker ML and KG signals; (right) incorrect prediction where mismatched RAG evidence introduces uncertainty despite ML and KG supporting direct binding.

the LLM overgeneralizes (e.g., inferring associations for EPHA2 from EPHA1) (Figure 3D).

These results indicate DrugAgent resolves conflict in a manner that preserves agreement and moderates uncertainty, while deferring to uncertainty when evidence conflicts.

5. Discussion

This study addresses a central challenge in the use of AI in science: how to aggregate heterogeneous evidence into interpretable decisions. Structuring diverse evidence sources remains challenging. DrugAgent addresses this problem by separating reasoning steps from deterministic decision-making to enable structured aggregation of ML, KG, and RAG. This design yields faithful and consistent decisions,

even under conflicting evidence. A key insight from our analysis is that plausibility does not imply better evidence alignment. In particular, removing constraints such as RAG can increase plausibility scores while degrading grounding, highlighting a gap between perceived reasoning quality and actual evidence consistency. This suggests that effective reasoning by LLMs for scientific applications remains challenged in scenarios of conflict and uncertainty.

6. Impact Statement

This paper presents DrugAgent, a framework for aggregating heterogeneous biomedical evidence into structured and interpretable decisions. It may support hypothesis generation and improve transparency in multi-source reasoning. However, outputs may appear plausible despite imperfect evidence alignment. The system should be used as a decision-support tool with human expert validation, not as a substitute for experimental or clinical judgment.

References

Agarwal, C., Tanneru, S. H., and Lakkaraju, H. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*, 2024.

Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H., and Peterson, J. R. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1039–1045, 2011.

Davis, A. P., Wiegers, T. C., Johnson, R. J., Sciaky, D., Wiegers, J., and Mattingly, C. J. Comparative toxicogenomics database (ctd): update 2023. *Nucleic acids research*, 51(D1):D1257–D1262, 2023.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Durmus, E., He, H., and Diab, M. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5055–5070, 2020.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

He, T., Heidemeyer, M., Ban, F., Cherkasov, A., and Ester, M. SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of cheminformatics*, 9(1):1–14, 2017.

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S. E. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, 6: e26726, 2017.

Huang, K., Fu, T., Glass, L. M., Zitnik, M., Xiao, C., and Sun, J. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23): 5545–5547, 2020.

Ioannidis, V. N., Song, X., Manchanda, S., Li, M., Pan, X., Zheng, D., Ning, X., Zeng, X., and Karypis, G. Drkg - drug repurposing knowledge graph for covid-19. https://github.com/gnn4dr/DRKG/, 2020.

James, F., Churas, C., Pratt, D., and Luna, A. texttoknowledgegraph: Generation of molecular interaction knowledge graphs using large language models for exploration in cytoscape. *bioRxiv*, 2025.

Ki, D., Rudinger, R., Zhou, T., and Carpuat, M. Multiple llm agents debate for equitable cultural alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24841–24877, 2025.

Knox, C., Wilson, M., Klinger, C. M., Franklin, M., Oler, E., Wilson, A., Pon, A., Cox, J., Chin, N. E., Strawbridge, S. A., et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research*, 52(D1):D1265–D1275, 2024.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Liu, T., Hwang, L., Burley, S. K., Nitsche, C. I., Southan, C., Walters, W. P., and Gilson, M. K. Bindingdb in 2024: a fair knowledgebase of protein-small molecule binding data. *Nucleic acids research*, 53(D1):D1633–D1644, 2025.

Lynch, C., Sakamuru, S., Huang, R., Stavreva, D. A., Varticovski, L., Hager, G. L., Judson, R. S., Houck, K. A., Kleinstreuer, N. C., Casey, W., et al. Identifying environmental chemicals as agonists of the androgen receptor by using a quantitative high-throughput screening platform. *Toxicology*, 385:48–58, 2017.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., and Callison-Burch, C. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of*

the Association for Computational Linguistics (Volume 1: Long Papers), pp. 305–329, 2023.

Öztürk, H., Özgür, A., and Ozkirimli, E. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Petridis, P., Margaritis, G., Stoumpou, V., and Bertsimas, D. Holistic ai in medicine; improved performance and explainability. *npj Digital Medicine*, 2026.

Pitre, P., Ramakrishnan, N., and Wang, X. Consensagent: Towards efficient and effective consensus in multi-agent llm interactions through sycophancy mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22112–22133, 2025.

Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., and Vosoughi, S. Judging the judges: A systematic study of position bias in llm-as-a-judge. doi: 10.48550. *arXiv preprint arXiv.2406.07791*, 2024.

Shu, D., Chen, T., Jin, M., Zhang, Y., Du, M., and Zhang, Y. Knowledge graph large language model (kg-llm) for link prediction. *arXiv preprint arXiv:2403.07311*, 2024.

Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., and Gerstein, M. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2023.

Wang, A., Cho, K., and Lewis, M. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5008–5020, 2020.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022b.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., and Wang, C. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

Yao, B., Shang, C., Du, W., He, J., Lian, R., Zhang, Y., Su, H., Swamy, S., and Qi, Y. Peacemaker or troublemaker: How sycophancy shapes multi-agent debate. *arXiv preprint arXiv:2509.23055*, 2025.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623, 2023.

A. Software and Data

The framework is implemented using Microsoft AutoGen (Wu et al., 2023). All agents use an OpenAI GPT-5.2 model via the Azure OpenAI API. Deterministic decoding (temperature = 0.0, seed = 42) is used for all components except the reasoning module.

B. Related Works

Despite advances across these areas, a key unresolved challenge is how to systematically aggregate heterogeneous evidence from multiple sources into a unified, structured, and interpretable reasoning framework.

Machine Learning in Drug Target Interaction Deep learning models such as DeepPurpose (Huang et al., 2020) predict drug-target interactions from molecular and protein representations. However, these approaches operate as single-modality predictors and do not support structured aggregation of heterogeneous evidence within a unified reasoning framework.

Knowledge Graphs for Integrative Analysis. Knowledge graphs (KGs) integrate diverse biomedical data and have been widely used for link prediction and drug discovery (Himmelstein et al., 2017; Ioannidis et al., 2020). However, existing approaches typically rely on graph embeddings or path-based reasoning and do not support structured aggregation of heterogeneous evidence across multiple sources.

Literature Search using LLMs Retrieval-augmented generation (RAG) enables LLMs to incorporate external knowledge from the literature (Lewis et al., 2020; James et al., 2025). However, existing approaches focus on retrieval and summarization, rather than structured aggregation of heterogeneous evidence across multiple sources.

LLMs with Reasoning Techniques such as Chain-of-Thought and ReAct enable LLMs to perform multi-step reasoning (Wei et al., 2022b; Yao et al., 2022). However, these approaches rely on implicit or unstructured reasoning and do not support structured aggregation and alignment of heterogeneous evidence.

Multi-Agent Systems Multi-agent systems enable coordination among specialized agents and have been widely adopted in LLM-based applications (Du et al., 2023; Tang et al., 2023). However, existing approaches lack principled mechanisms for structured aggregation of heterogeneous evidence across agents.

Consensus and Decision-Making in Multi-Agent LLMs Recent multi-agent LLM systems often rely on consensus mechanisms such as voting or debate to improve reasoning (Ki et al., 2025). While these approaches leverage diverse agent perspectives, they suffer from key limitations, including sycophancy and premature consensus under conflicting evidence (Pitre et al., 2025; Yao et al., 2025). Such dynamics can suppress meaningful disagreement and degrade reasoning reliability, indicating that agreement alone is insufficient and that explicit modeling of conflict and uncertainty is necessary for principled aggregation.

Collectively, existing approaches address individual components—prediction, retrieval, reasoning, and coordination—but lack a principled mechanism to reconcile conflicting evidence across sources. This gap motivates our structured aggregation framework.

C. Datasets

C.1. Kinase Inhibition Benchmark

For quantitative evaluation, we used a kinase-compound activity dataset from a large-scale profiling study (Anastassiadis et al., 2011). The dataset reports kinase activity as the percentage of remaining enzymatic function after compound exposure and includes 300 human kinases and 178 small-molecule inhibitors.

Drug identifiers were standardized to canonical representations. Activity values were normalized to $[0, 100]\%$ and discretized into three classes based on remaining activity: Strong ($\leq 25\%$), Moderate (25-50%), and Weak (50-80%), following commonly used thresholds in kinase profiling studies (Anastassiadis et al., 2011).

From 1,895 labeled interactions, we constructed a balanced benchmark of 900 pairs (300 per class), covering 178 kinases

and 42 inhibitors. Sampling was performed under diversity constraints, and interactions were stratified by knowledge graph support.

C.2. Tox21 AR Dataset

The Tox21 androgen receptor (AR) dataset is derived from the Tox21 10K screening initiative and consists of compound-target activity annotations from cell-based assays (Lynch et al., 2017), with labels indicating Active or Inactive compounds.

We constructed an AR-specific subset by filtering for consistent assay conditions and restricting to compounds present in a knowledge graph. Compound-target pairs were ranked based on supporting literature evidence, and the top 250 active and 250 inactive samples were selected to form a balanced dataset of 500 pairs.

Table 1. Summary of evaluation datasets. The kinase dataset provides a controlled, class-balanced benchmark, while the Tox21 AR dataset represents a more realistic and challenging setting for evaluating aggregation under noisy and heterogeneous evidence.

	Kinase	Tox21 (AR)
Total pairs	900	500
Class distribution	300 / 300 / 300	250 / 250
Class labels	Weak / Moderate / Strong	Active / Inactive
Unique drugs	42	500
Unique targets	178	1 (Androgen Receptor)
Assay type	Enzymatic inhibition	Cell-based reporter assay
Prediction task	Multi-class	Binary

D. Experimental Setup

D.1. Evaluation Framework

Evaluation Objective. The primary objective of this study is to evaluate DrugAgent as a framework for faithful multi-tool evidence aggregation under an LLM-as-a-Judge paradigm. Rather than focusing on predictive performance, we assess whether structured integration of heterogeneous signals produces evidence-aligned and internally consistent conclusions.

As a secondary validation, we report three-class DTI classification results to verify alignment with experimental labels.

Datasets and Evaluation Regimes. Evaluation is conducted on a kinase benchmark dataset, with an additional Tox21 dataset used for complementary analysis under differing statistical and biological characteristics.

The kinase benchmark provides a controlled, class-balanced enzymatic inhibition setting with well-characterized quantitative binding measurements and serves as the primary testbed for assessing aggregation fidelity under stable conditions.

In contrast, the Tox21 screening dataset reflects a realistic high-throughput assay environment, characterized by limited structured evidence and noisier measurements. Rather than serving as a primary benchmark for predictive performance, it is used to examine aggregation behavior under more challenging, real-world conditions.

Together, these datasets enable a systematic analysis of aggregation fidelity across both controlled experimental settings and biologically realistic scenarios.

Ablation Setting. To evaluate the contribution of each component, we perform a full ablation study in which each module (ML, KG, RAG) is explicitly removed from the pipeline. In each ablation setting, the system is re-executed using only the remaining sources, and the removed component is excluded from both the reasoning stage and the final decision process. This ensures that no information from the ablated source is used during aggregation, enabling a direct assessment of each component's contribution.

D.2. Aggregation Fidelity under LLM-as-a-Judge

Aggregation quality is evaluated using an LLM-as-a-Judge protocol grounded in recent analyses of evaluation reliability, bias, and faithfulness (Gu et al., 2024; Zheng et al., 2023; Shi et al., 2024).

We operationalize aggregation fidelity along four complementary dimensions:

(1) Faithfulness. Following prior work on faithfulness and attribution in LLM evaluation (Agarwal et al., 2024; Lyu et al., 2023; Gu et al., 2024), we assess whether the generated fusion explanation is grounded in the provided evidence and remains consistent with it. The judge evaluates the explanation in relation to the available ML evidence, KG evidence, and retrieved literature evidence, and checks whether the response introduces unsupported claims, contains contradictions, or omits critical information necessary to justify the conclusion. Faithfulness is scored on a five-point scale: 1 = highly unfaithful, 2 = mostly unfaithful, 3 = mixed/uncertain, 4 = mostly faithful, and 5 = fully faithful. In addition to the overall score, the judge also outputs structured binary indicators indicating whether the explanation is grounded in the input, contains contradictions, introduces unsupported claims, or omits critical information, along with a brief rationale. This formulation is related to QA-based factual consistency metrics such as FEQA (Durmus et al., 2020) and QAGS (Wang et al., 2020).

(2) Biological Plausibility. Biological plausibility evaluates whether the inferred drug-target association constitutes a reasonable biological hypothesis given the available evidence (Petridis et al., 2026).

Specifically, the judge assesses (i) whether the input evidence contradicts the association and (ii) whether the reasoning is mechanistically coherent. Plausible associations are those that are not contradicted and are supported by biologically consistent mechanisms, even in the absence of direct experimental confirmation.

The judge assigns a 5-point Likert score (1: Strongly Disagree to 5: Strongly Agree), where lower scores reflect contradiction or weak biological grounding, and higher scores indicate mechanistically coherent and evidence-consistent hypotheses. In addition to the score, the judge outputs binary indicators for contradiction and mechanistic coherence, along with a brief rationale.

(3) Stability. Stability evaluates the reproducibility of aggregation outcomes under repeated executions. We randomly sample $N = 100$ cases and run the aggregation three times ($k = 3$).

We report **label stability** as the fraction of cases for which all runs yield identical categorical outputs:

$$\text{AgreementRate} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i^{(1)} = \hat{y}_i^{(2)} = \hat{y}_i^{(3)}\}.$$

Higher AgreementRate indicates greater robustness to stochastic variation.

(4) Evidence Combination Consistency and Conflict Dynamics. We analyze aggregation behavior conditioned on combinations of module outputs (ML, KG, RAG), representing each case as a triplet (e.g., W—M—S where W is weak, M is moderate, and S is strong). For each combination, we compare the distribution of ground-truth labels and aggregated outputs to assess consistency under similar evidence configurations.

We further examine label transitions across the pipeline, from module predictions to the final decision, to characterize how conflicting signals are resolved and integrated.

D.3. Sanity-Check DTI Classification

To quantify predictive consistency, we compute Accuracy and Macro-F1 as standard classification metrics. We compare individual modules (ML, KG, RAG), ablation variants (No ML, No KG, No RAG), and the full DrugAgent model.

Leakage Control. To prevent label leakage, ground-truth interaction labels (binding_class) were not exposed to any agent during aggregation. Retrieval was restricted to PMC Open Access articles and curated knowledge graph sources, excluding evaluation annotations. The LLM judge was provided only with raw module outputs and final explanations, ensuring independence from ground-truth labels.

E. Secondary Results

E.1. Aggregation Fidelity

(5) Evaluation of Generalizability across Large-scale Chemical Spaces To evaluate generalization beyond the kinase setting, we conduct experiments on the Tox21 AR dataset (Fig. 4), which reflects a more diverse and realistic chemical space.

Figure 4A shows that DrugAgent maintains a high proportion of "Good" rationales (34.4%), while removing KG leads to a substantial increase in "Plausible only" outputs (51.2%), indicating reduced mechanistic grounding.

Figure 4B further shows that removing RAG shifts plausibility scores upward (mean $\simeq 2.88$) but also increases the proportion of contradictory "Bad" cases (24.6%), suggesting that external evidence acts as a constraint that prevents overconfident but unsupported reasoning.

Figure 4C demonstrates that DrugAgent better separates Active from Inactive compounds, with higher median scores for active samples. This separation degrades when ML or KG is removed, indicating that both molecular signals and structured knowledge are necessary for effective generalization.

Together, these results show that KG, ML, and RAG play complementary roles: KG supports mechanistic reasoning, ML provides discriminative signals, and RAG enforces evidence grounding. Their integration enables robust generalization across diverse chemical spaces.

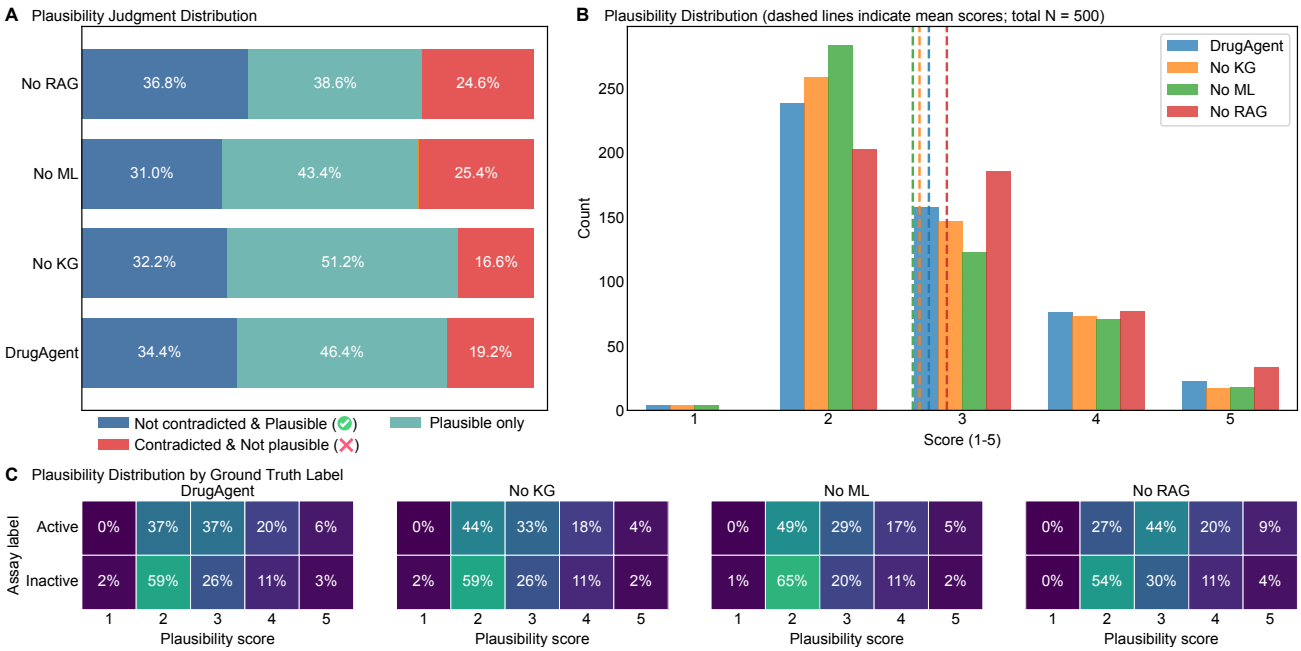

Figure 4. **Generalizability analysis and statistical validation using the Tox21 benchmark.** (A) Distribution of logical consistency states across the Tox21 AR dataset, highlighting the role of KG and RAG in maintaining reasoning quality in diverse chemical spaces. (B) Kernel density estimation (KDE) of plausibility scores, illustrating the impact of component ablation on the model's confidence distribution. (C) Assessment of bio-activity discrimination. The split violin plot shows that DrugAgent's integrated reasoning consistently separates Active (blue) from Inactive (gray) compounds, validating its utility as a generalizable metric for drug activity prioritization .

E.2. Sanity-Check DTI Classification

As a complementary validation, we evaluate whether the aggregation mechanism preserves alignment with experimentally derived interaction labels in a three-class DTI classification setting. This analysis is not intended to optimize predictive performance, but to verify that structured aggregation remains consistent with empirical labels.

As shown in Table 2, combining heterogeneous sources improves performance over standalone models, with simple combinations such as ML+KG providing strong baselines.

Category	Model	Kinase		Tox21	
		Acc	Macro-F1	Acc	Macro-F1
Standalone	ML	0.372	0.370	0.612	0.608
	KG	0.380	0.359	0.582	0.549
	RAG	0.341	0.338	0.526	0.485
Naive Combinations	ML + KG	**0.408**	**0.407**	**0.632**	**0.623**
	ML + RAG	0.383	0.373	0.596	0.585
	KG + RAG	0.376	0.364	0.564	0.547
Structured	DrugAgent	0.391	0.390	0.590	0.569

Table 2. Full ablation across standalone models, pairwise combinations, and structured aggregation (DrugAgent). While naive combinations (especially ML+KG) achieve the highest accuracy, DrugAgent provides competitive performance with structured multi-source reasoning, as analyzed further in subsequent sections. Abbreviation: Accuracy (Acc)

DrugAgent achieves competitive performance relative to these baselines, while maintaining alignment with ground-truth labels despite not being optimized for classification accuracy.

E.3. Aggregation Fidelity

(1) Faithfulness. Faithfulness scores are consistently high across all configurations, with most samples rated 4 or 5 (Agree or Strongly Agree), indicating strong alignment between generated explanations and the provided evidence. This is expected, as all inputs and outputs are explicitly grounded in the underlying evidence sources by design. The number of evaluated samples is comparable across configurations (Full: $n = 889$, No ML: $n = 896$, No KG: $n = 857$, No RAG: $n = 900$). While high faithfulness reflects consistent grounding, it does not necessarily indicate correct or biologically meaningful conclusions.

(3) Stability. To evaluate reproducibility, we conducted a stability analysis on a randomly sampled subset of $N = 100$ drug-target pairs, running each case three times ($k = 3$).

Label stability, defined as the fraction of cases with identical outputs across runs, shows a mean agreement of 0.983, a median of 1.000, and a perfect agreement rate of 0.95, indicating near-deterministic behavior with only minor variability across runs.

Reasoning stability, measured via embedding-based semantic similarity between explanations, yields a mean similarity of 0.881, a median of 0.884, and a minimum of 0.760, indicating that explanations remain highly consistent despite minor variations in phrasing.

Together, these results demonstrate that the aggregation process is both label-stable and reasoning-consistent, supporting the reproducibility of the framework.

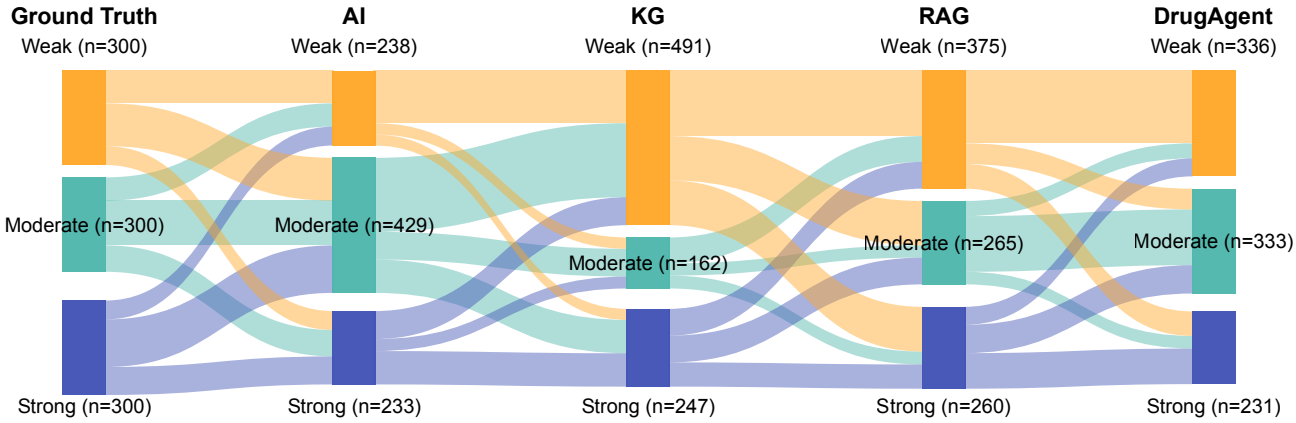

Figure 5. Flow of predictions from ground truth through ML, KG, and RAG modules to the final DrugAgent output. Each stage shows the distribution of labels (Strong, Moderate, Weak) and how evidence is progressively integrated.

To analyze how heterogeneous evidence is integrated, we examine aggregation behavior conditioned on combinations of module outputs (ML, KG, RAG), representing each case as a triplet (e.g., M—W—S). We compare ground-truth labels and final outputs across these patterns, and analyze how predictions evolve through the aggregation pipeline.

Figure 5 shows that individual modules exhibit distinct biases, with ML favoring Moderate predictions, KG skewing toward Weak labels, and RAG introducing higher variability. These biases lead to distorted label distributions when considered in isolation. In contrast, DrugAgent mitigates these effects by rebalancing the final outputs while preserving strong consensus signals. When modules agree, predictions are largely retained, whereas conflicting signals are consistently mapped to Moderate, reflecting uncertainty-aware aggregation. This behavior demonstrates that DrugAgent both corrects systematic biases and enforces structured reconciliation across heterogeneous evidence sources.

F. Method Details

F.1. Overview of DrugAgent

DrugAgent is a multi-agent architecture for biologically grounded assessment of drug–target interactions. The system integrates complementary evidence sources—machine learning predictions, knowledge graph reasoning, and literature-derived signals—within a structured aggregation pipeline.

A key design principle is the separation of the reasoning step from deterministic decision-making. Evidence from heterogeneous sources is first organized into a structured representation and then reconciled through a rule-guided decision module, enabling consistent and interpretable outputs.

As illustrated in Fig. 1, the system follows a modular pipeline. A coordination layer manages execution, while specialized agents independently generate complementary evidence: an ML Agent produces quantitative affinity predictions, a KG agent extracts mechanistic paths, and a RAG agent retrieves literature evidence. These signals are integrated by a reasoning module that produces the final aggregated prediction and explanation.

Pipeline. The system operates in three stages: (1) coordination, which orchestrates agent execution in either a fast deterministic mode or via an optional coordinator agent; (2) evidence generation, where each agent produces a label and supporting rationale; and (3) aggregation, where a reasoning module synthesizes evidence and applies rule-guided fusion to produce the final decision.

Implementation. The framework is implemented using Microsoft AutoGen (Wu et al., 2023). All agents use an OpenAI GPT-5.2 model via the Azure OpenAI API. Deterministic decoding (temperature = 0.0, seed = 42) is used for all components except the reasoning module. The RAG pipeline uses the `text-embedding-3-large` model for retrieval. Additional implementation details are provided in the following sections.

F.1.1. Coordination Layer and Optional Coordinator Agent

The coordination layer orchestrates the overall workflow by invoking specialized agents either through direct function calls (fast mode) or via an optional Coordinator Agent. In the default fast mode, modules are executed independently without iterative interaction, ensuring efficiency and reproducibility.

When enabled, the Coordinator Agent decomposes tasks and manages execution through structured coordination across agents.

The coordination layer also enforces standardized JSON schemas, ensuring consistent and structured outputs across all agents.

F.1.2. ML Agent

The ML Agent predicts drug–target interactions using DeepPurpose (Huang et al., 2020), a pre-trained model on BindingDB (Liu et al., 2025) that estimates continuous pK_d values from molecular SMILES strings and protein sequences.

For the kinase benchmark, predicted affinities are discretized into three ordinal categories following standard conventions: Strong ($pK_d \geq 7.0$), Moderate ($6.0 \leq pK_d < 7.0$), and Weak ($pK_d < 6.0$) (Öztürk et al., 2018; He et al., 2017).

For the Tox21 dataset, predictions are mapped to binary labels, where Strong is treated as Active, and Moderate or Weak as

Inactive, consistent with assay definitions.

F.1.3. KG AGENT

We construct a heterogeneous Drug–Gene knowledge graph by integrating curated biomedical resources, including DrugBank (Knox et al., 2024), CTD (Davis et al., 2023), and BindingDB (Liu et al., 2025). Nodes represent drugs and genes, and edges encode curated interactions with associated evidence.

For each Drug–Gene pair, textual evidence from these sources is aggregated into structured, source-aware representations, preserving interpretability across curated annotations and experimental measurements.

We perform multi-hop reasoning over this graph using a breadth-first search with a maximum depth of $H = 5$. Candidate paths are ranked by a structural scoring function that favors shorter paths while penalizing high-degree hub nodes:

$$S(P) = \frac{1}{|P|} \cdot \lambda^{H(P)},$$

where $|P|$ is the path length and $H(P)$ counts hub nodes (degree > 300). We use $\lambda = 0.6$ and select the top-k paths ($k = 3$) for downstream reasoning. These hyperparameters were selected based on sensitivity analysis, demonstrating stable performance across a reasonable range of values.

The selected paths are processed by an LLM-based component to produce a structured output consisting of a categorical label, confidence score, and supporting rationale. This design enables interpretable, mechanism-aware reasoning while reducing the influence of non-specific graph connectivity.

Compared to embedding-based link prediction methods, this path-based approach preserves explicit mechanistic interpretability, which is critical for downstream evidence aggregation.

F.1.4. PUBMED RAG AGENT

We construct an unstructured evidence layer using a retrieval-augmented generation (RAG) pipeline over PubMed Central (PMC) Open Access articles. To address sparse direct co-mentions, we employ a multi-channel retrieval strategy that issues queries at three levels: drug–target pairs, drug-only, and target-only. For each query type, we retrieve up to 10 documents per entity type to ensure balanced coverage across sources.

As summarized in Table 3, this strategy provides broad coverage across both kinase and Tox21 domains despite their differing data distributions, demonstrating the effectiveness of multi-channel retrieval.

	Kinase	Tox21 (AR)
Article Statistics		
Total unique PMC full-text articles	2,765	6,743
Retrieval Channels		
Pair-level (co-mention) articles	722	795
Drug-level (single-entity) articles	333	4,223
Gene-level (single-entity) articles	1,710	10
Entity Coverage		
Pairs with $\geq$1 co-mention article	178	413
Drugs with $\geq$1 article	39	491
Genes with $\geq$1 article	178	1

Table 3. Summary statistics for the kinase and Tox21 (AR) datasets, including article counts across retrieval channels and entity-level coverage in the multi-channel RAG pipeline.

Retrieval and Scoring Query embeddings are generated using the `text-embedding-3-large` model, with assay-related terms appended to prioritize experimentally relevant content. Retrieved passages are re-ranked using a heuristic scoring scheme that combines embedding similarity with structural signals, favoring direct interaction evidence and filtering low-information fragments.

Validation To ensure grounding, retrieved evidence is validated through a rule-based pipeline. Outputs must be supported by retrieved text, explicitly reference both query entities, and rely on either direct co-mention evidence or consistent multi-source support. Otherwise, confidence is downgraded.

All retrieved evidence is preserved with provenance metadata, enabling traceability to source articles and text spans.

F.1.5. REASONING AGENT

The Reasoning Agent operates in two stages: structured reasoning and rule-guided decision-making.

In the first stage, outputs from the ML, KG, and RAG modules are converted into a structured JSON format that allows communication reasoning logic between sub-agents. Each source contributes an *evidence_analysis* entry that includes evidence source, agent thoughts, proposed action, and observations about the input, along with its proposed DTI label. A summarization step then adds a *summary_reasoning* field that captures agreement, conflict, and uncertainty across sources. This stage summarizes evidence but does not produce the final decision.

In the second stage, a decision module consumes the structured reasoning and produces the final output, including a label (*Weak*, *Moderate*, *Strong*) and supporting rationale. Decisions follow a rule priority order: majority agreement is applied first, followed by tie-breaking described in the prompt using KG structure, pK_d thresholds, and explicit evidence from RAG (in that order). When these rules are inconclusive, a constrained LLM-based fallback is applied as a last resort.

This separation between reasoning and rule-guided decision-making ensures stable and reproducible outputs while preserving expressive intermediate reasoning.