

UniD³: Open Knowledge Graphs and QA Benchmarks for Drug–Disease Discovery and Reasoning

Qing Wang¹, Tianshi Liu², Minghao Zhou³, Jialu Liang¹,
Sen Guo³, Jing Su³, Qianqian Song^{4, #}

¹Department of Health Outcomes and Biomedical Informatics, University of Florida

²H. Lee Moffitt Cancer Center and Research Institute

³Department of Biostatistics and Health Data Science, Indiana University School of Medicine

⁴School of Applied and Creative Computing, Purdue University

[#]Correspondence: song1081@purdue.edu

Abstract

Drug–disease knowledge underpins drug discovery and precision medicine, yet most of it is locked in unstructured PubMed articles and only partially captured by curated databases. We release UniD³-KG, a suite of six knowledge graphs spanning three tasks (Drug–Disease Matching, DDM; Drug Effectiveness Assessment, DEA; and Drug–Target Analysis, DTA), and UniD³-Data, three open QA datasets (28,915 DDM, 15,042 DEA and > 4,000 DTA pairs), distilled from 157,849 PubMed articles with Llama-3.3-70B by the UniD³ KG-RAG pipeline, which couples task-specific prompting with a *dual-stage entity extraction* that re-extracts over the consolidated graph to recover cross-chunk and cross-document evidence. External validation on CDR, DrugReviews, P3ps and MedicationQA and a benchmark across seven LLM baselines and three retrieval variants show that UniD³-Data is faithful to curated sources and that retrieval over UniD³-KG consistently improves biomedical QA. Validation strength is uneven by design and we report it as such: DDM and DEA are checked at the triple level ($F1 = 0.85\text{--}0.87$) while DTA is checked only at the answer level. Three clinicians independently re-rated a 469-drug, 2,652-pair sample of the DDM split, agreeing almost perfectly with each other (Fleiss' $\kappa = 0.954$) and with UniD³-Data at majority-vote accuracy 0.900. All artefacts—UniD³-KG, UniD³-Data, benchmark scripts and the UniD³-QA explainable chatbot—are released¹ as shared infrastructure for biomedical NLP and drug repurposing.

1 Introduction

Systematic characterisation of drug–disease relationships underpins drug discovery, repurposing, and precision medicine (Yang et al., 2014; Corrigan-Curay et al., 2018; Liu et al., 2024). Yet the bulk of this knowledge is buried in millions

of unstructured biomedical articles, while curated databases such as DrugBank (Wishart et al., 2018) or Open Targets (Ochoa et al., 2021) cover only a small fraction of the literature (Zhu et al., 2019). Closing this gap requires NLP systems that can reliably surface drug, disease and target entities *and* the relations among them from heterogeneous biomedical text.

Most existing biomedical relation extraction pipelines rely on manual curation or on rule-based and supervised models trained on small annotated corpora (Lai et al., 2023; Davis et al., 2009). While accurate within their training distribution, such pipelines suffer from limited coverage, slow update cycles and brittle generalisation across drug families, disease ontologies and writing styles. Recent open biomedical LLMs (BioGPT, Llama3-OpenBioLLM, MedLlama3) close part of the gap on question answering, but, used directly, they remain prone to hallucination and offer no structured artefact that downstream biomedical applications can audit or query (Luo et al., 2022; Pal and Sankarasubbu, 2024; Chen et al., 2024).

Retrieval-augmented generation (RAG), and in particular knowledge-graph augmented variants (Guo et al., 2025; Edge et al., 2024; Han et al., 2024; Matsumoto et al., 2024; Wu et al., 2025; Li et al., 2024, 2025), has emerged as an attractive alternative. An LLM is conditioned on retrieved evidence and on an explicit graph, which both grounds generation and exposes a structured by-product. However, off-the-shelf KG-RAG systems are not biomedical-aware. They use generic extraction prompts, perform single-pass extraction that misses cross-chunk and cross-document evidence, and are evaluated on general-domain corpora. As a result there is, to our knowledge, no large-scale, openly available drug–disease KG-RAG resource that NLP and biomedical-informatics researchers can reuse and benchmark against.

We introduce UniD³, an application-oriented

¹<https://github.com/QSong-github/UniD3>

framework that adapts KG-RAG to drug-centric biomedical literature. The central contribution is a bundle of openly released biomedical artefacts; the methodological design (task-specific prompting and dual-stage extraction) is what makes them possible, and we do not claim it as a major algorithmic advance. Throughout we separate two kinds of number: benchmark scores computed against UniD³-KG-derived references measure internal consistency and are marked as upper bounds, whereas independently validated biomedical correctness comes only from the external benchmarks and the clinician study. Concretely, UniD³ contributes:

- *UniD³-KG* — *externally validated open knowledge graphs* distilled from 157,849 PubMed articles: six graphs spanning DDM, DEA and DTA ($\sim 112\text{K}$ entities, $\sim 300\text{K}$ evidence-grounded edges after Stage II; Table 2). Triple-level F1 reaches 0.85–0.87 against the CDR, DrugReviews and P3ps external benchmarks.
- *UniD³-Data* — *three task-specific QA datasets* (28,915 DDM, 15,042 DEA and $> 4,000$ DTA pairs), each carrying a supporting KG path π for auditability. The DDM split agrees with a three-clinician verification on 469 FDA-approved drugs at majority-vote accuracy 0.90, Fleiss’ $\kappa = 0.95$ (Section 5.3).
- *A method bundle*: task-specific prompts with cross-LLM agreement filtering and a dual-stage extraction procedure that recovers cross-document evidence.
- *Released infrastructure for downstream use*: a systematic benchmark of seven biomedical and general LLMs and three LightRAG-augmented variants with reproducible scripts and per-seed scores, and UniD³-QA, an explainable chatbot that returns answers paired with the supporting KG path and PubMed evidence sentences.

2 Related Work

Biomedical text mining and relation extraction.

Earlier systems for drug, disease and target extraction rely on hand-crafted rules, ontology-aware NER, supervised relation extractors trained on heterogeneous corpora (CTD, BioREx), similarity-constrained matrix factorisation over curated association matrices, and aggregated curated KGs

such as Hetionet (Himmelstein et al., 2017) and ROBOKOP (Morton et al., 2019) (Davis et al., 2009; Lai et al., 2023; Zhu et al., 2019; Yildirim et al., 2007; Tanoli et al., 2018; Zhang et al., 2018). These resources offer high relational precision but generalise poorly across drug classes and disease ontologies, require costly re-training to keep pace with the literature, and rarely carry sentence-level provenance. UniD³ is complementary. It trades some per-relation curation for prompted LLM extraction conditioned on task-specific schemas, retaining explicit PubMed evidence sentences at substantially higher coverage and an order of magnitude lower curation cost.

LLM-based dataset construction. LLM-driven curation has been used to augment tabular benchmarks (Seedat et al., 2024), generate synthetic dialogue corpora (Abdullin et al., 2023) and unify drug toxicity labels from heterogeneous FDA documents (Silberg et al., 2024). These efforts show that LLM outputs, when filtered, can match or complement expert-curated resources. We extend this line to drug–disease discovery. UniD³ produces *structured* resources (graphs and QA datasets) rather than free-text augmentation, and grounds them on KG paths so that downstream consumers can audit each generation.

Biomedical question answering and benchmarks.

Domain-specialised biomedical LLMs such as BioGPT (Luo et al., 2022), MedLlama3 (John Snow Labs, 2024), MedX (skumar9, 2024), HuatuoGPT-o1 (Chen et al., 2024) and Llama3-OpenBioLLM (Pal and Sankarasubbu, 2024) push the state of the art on medical QA, but they are evaluated almost exclusively on closed benchmarks (PubMedQA (Jin et al., 2019), MedQA, MedicationQA) whose triple-level evidence is not openly available. Our benchmark complements this line in three respects. The released datasets cover three task-specific axes at scale. Both structured metrics (Cov@ k , F1, AUROC) and free-form metrics (BLEU, ROUGE, BERTScore) are reported on the *same* backbone models, allowing direct comparison between extractive coverage and generative faithfulness. The underlying KG paths are released alongside the QA pairs, so that any disagreement between predicted and benchmark answers is traceable to specific PubMed evidence rather than to opaque model behaviour. A complementary benchmark grounds its evidence in curated regulatory records instead of literature: Wang et al. (2026)

score whether a cited source is the upstream primary record, where our items trace to PubMed passages.

Retrieval-augmented generation with knowledge graphs. GraphRAG-style systems combine retrieval over a constructed graph with generative reasoning, including general-purpose engines such as LightRAG (Guo et al., 2025), MiniRAG (Fan et al., 2025) and Microsoft GraphRAG (Edge et al., 2024; Han et al., 2024), as well as biomedical adaptations such as BiomedRAG (Li et al., 2025), Med-GraphRAG (Wu et al., 2025), DALK (Li et al., 2024) and KRAGEN (Matsumoto et al., 2024). Concurrent 2025–2026 work explores two-stage abstract-then-fulltext KG construction (Meng et al., 2025) and multi-LLM consensus filtering for clinical KGs (Das et al., 2026). Most biomedical KG-RAG systems still target a single condition (e.g. Alzheimer’s) or a single benchmark (e.g. MedQA), use single-pass extraction, and do not release large-scale public training/evaluation data. UniD³ extends this line of work along three dimensions: it scales the corpus to 157,849 PubMed articles spanning three biomedical tasks rather than a single condition or benchmark; it adopts a dual-stage extraction strategy that aggregates evidence across chunks and documents instead of relying on a single forward pass; and the resulting graphs, datasets, benchmarks and chatbot are released as shared infrastructure for biomedical NLP.

3 The UniD³ Framework

3.1 Overview and notation

Let $\mathcal{D} = \{d_1, \dots, d_N\}$ denote a corpus of biomedical articles retrieved from PubMed and let $\mathcal{T} = \{\text{DDM}, \text{DEA}, \text{DTA}\}$ index the three target tasks. UniD³ produces, for every task $t \in \mathcal{T}$, a knowledge graph $\mathcal{G}^{(t)} = (\mathcal{V}^{(t)}, \mathcal{E}^{(t)}, \mathcal{R}^{(t)})$ over biomedical entities $\mathcal{V}^{(t)}$, typed relations $\mathcal{R}^{(t)}$ and sentence-level evidence $\mathcal{E}^{(t)}$, together with a task-specific question–answer dataset whose tuples (q, a, π) pair every query q with a reference answer a and the supporting KG path π . The pipeline (Figure 1) is organised into four sequential stages: a task-specific classifier first routes $\mathcal{D}$ into the per-task corpora $\mathcal{D}_{\text{src}}^{(t)}$; a prompt-construction step then assembles the entity- and relation-extraction prompt $\mathcal{P}_t$; a dual-stage extraction procedure yields the knowledge graph $\mathcal{G}^{(t)}$; and a final KG-RAG step produces the per-task datasets and answers. We in-

stantiate the LLM throughout with Llama-3.3-70B (Grattafiori et al., 2024), denoted $\text{LLM}_\theta(\cdot)$.

3.2 PubMed retrieval and task-specific classification

We query PubMed with the keyword `Drug` and retrieve all available articles up to April 2025, yielding $|\mathcal{D}| = 157,849$ articles after de-duplication and language filtering (Appendix B). Each article is then routed by LLM_θ , under a task-specific classification prompt $\mathcal{P}_{\text{cls}}$, to one of $\{\text{DDM}, \text{DEA}, \text{DTA}\}$ or to an “other” bucket $\emptyset$ for off-topic articles. Articles in $\emptyset$ are discarded; the rest form the per-task corpora $\mathcal{D}_{\text{src}}^{(t)}$, with sizes 3,608 (DDM), 3,512 (DEA) and 3,688 (DTA). The full prompt template is given in Appendix A.

3.3 Task-specific extraction prompts

For each task t we hand-design an extraction prompt $\mathcal{P}_t$ following common prompt-engineering patterns (White et al., 2023). Each prompt concatenates three blocks: an instruction $\mathcal{I}_t$ stating the extraction objective in biomedical terms; a typed schema $\mathcal{S}_t$ that specifies the admissible entity types (*drug, disease, biomarker, target, mechanism, outcome*) and relation types; and a few-shot block $\mathcal{F}_t$ of curated exemplars. We seed $\mathcal{F}_t$ with three independently written exemplars per task and retain only those whose extractions are consistent across four LLMs (LLM_θ , GPT-5.4, Claude-4.6-Sonnet, Gemini-3-Pro), controlling for prompt-induced bias. Prompt templates are built on top of LightRAG (Guo et al., 2025) and are listed in full in Appendix A.

3.4 Dual-stage entity extraction

A single forward pass over individual chunks misses entity co-occurrences that span chunks or documents, a recurrent issue in biomedical literature where mechanism, drug and disease are often discussed in different sections of an article. We address this with a two-stage extraction procedure.

Stage I: paper-level extraction. Each article $d_j \in \mathcal{D}_{\text{src}}^{(t)}$ is segmented into overlapping chunks $\{c_{j,1}, \dots, c_{j,m_j}\}$, and the LLM extracts a set of typed triples from each chunk under $\mathcal{P}_t$:

$$\text{Tri}(c) = \text{LLM}_\theta(c; \mathcal{P}_t), \quad (1)$$

where $\text{Tri}(c) = \{(h, r, o)\}$. Triples are aggregated over the article and deduplicated by merging string

UniD3: Automated Drug–Disease Relationship Structuring

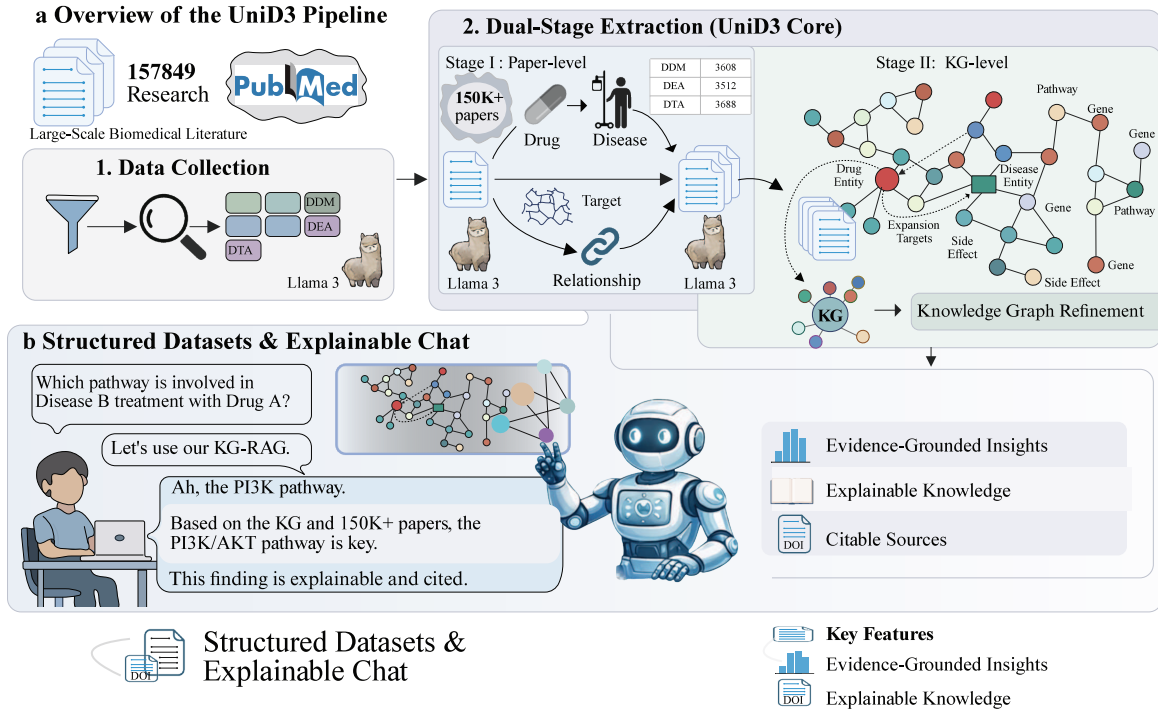


Figure 1: Overview of the UniD³ framework. (a) Drug-related PubMed articles are retrieved and routed by a task-specific classifier (Section 3.2) into the DDM, DEA and DTA corpora, after which the dual-stage entity extraction procedure (Section 3.4) produces six knowledge graphs and three open datasets via Llama-3.3-70B inside the LightRAG framework. (b) The constructed graphs are served to an explainable biomedical QA chatbot in which every answer is paired with the supporting KG path π and the underlying PubMed evidence sentences.

variants of the same entity whenever their normalised Levenshtein similarity exceeds a threshold τ , yielding the Stage I graph $\mathcal{G}_1^{(t)}$.

Stage II: KG-level re-extraction. For every entity v in $\mathcal{G}_1^{(t)}$ we serialise its h -hop neighbourhood, together with the sentence-level evidence attached to each edge, into a *KG-level chunk* c_v^{KG} , and apply Eq. 1 a second time with $c = c_v^{KG}$. The resulting triples are merged with $\mathcal{G}_1^{(t)}$ under the same threshold τ to form the Stage II graph $\mathcal{G}^{(t)}$. This second pass surfaces relations that connect entities across articles (e.g. *cisplatin* $\rightarrow$ *ferroptosis* $\rightarrow$ *ovarian cancer*) which Stage I alone cannot recover. Empirically (Section 5.1) Stage II increases edge count by $\Delta\mathcal{E}^{(t)}$ while keeping graph density $\rho(\mathcal{G}) = 2|\mathcal{E}|/(|\mathcal{V}|(|\mathcal{V}| - 1))$ in a regime that remains tractable for retrieval.

3.5 KG-RAG dataset and answer generation

Given $\mathcal{G}^{(t)}$, we instantiate dataset and answer generation inside LightRAG (Guo et al., 2025). LightRAG exposes four query modes over the constructed graph: *local* (entity-anchored neighbour-

hood retrieval), *global* (community-level retrieval), *naive* (chunk-level vector search over the source PubMed text), and *mix*, which integrates the entity- and community-level signals with the chunk-level vector evidence in a single retrieval call. We report all four for completeness and use LightRAG’s *mix* mode as the default setting; given a query q , the answer is generated by LLM $_{\theta}$ conditioned on q , the *mix*-mode retrieval over $\mathcal{G}^{(t)}$, and a task-specific generation prompt $\mathcal{P}_t^{\text{gen}}$.

Per-task instantiations. DDM samples (drug, disease) entity pairs from $\mathcal{G}^{(\text{DDM})}$, asks “which diseases is drug x matched to?” and stores the supporting KG path as π , yielding 28,915 pairs. DEA conditions on a (drug, indication) query and asks the model to rate effectiveness on a four-level scale, producing 15,042 instances. DTA formulates open-ended drug–target rationales with reference answers distilled from supporting evidence, giving $> 4,000$ QA pairs. All generations carry the retrieved KG path π so that downstream users can trace any answer back to its biomedical evidence (and audit the chatbot in Section 5.9).

4 Evaluation Setup

4.1 Datasets and tasks

We evaluate UniD³ on its three internal datasets and on four public biomedical benchmarks: CDR (chemical–disease relations from PubMed) (Li et al., 2016), DrugReviews (patient-reported drug indications and effectiveness ratings) (Gräßer et al., 2018), P3ps (condition-to-drug mappings) (P3ps, 2024) and MedicationQA (Abacha et al., 2019). CDR, DrugReviews and P3ps natively provide DDM-style drug–disease pairs; for the DEA evaluation we project gold-standard indication relations (CDR TREATS, DrugReviews positive ratings, P3ps listed mappings) to binary effective/ineffective labels, with absent or adverse-only relations treated as ineffective. This projection is deliberately coarse-grained: “ineffective”, “contraindicated”, “adverse”, “no evidence” and “not present in the benchmark” are clinically distinct states that the source benchmarks do not separate uniformly, so a finer mapping would not be recoverable from them. We therefore adopt the coarse rule and test how much the DEA result depends on where the line falls (§5.3). MedicationQA probes free-form drug QA aligned with DTA.

4.2 Reference models

We compare against seven biomedical and general LLMs without retrieval: Llama3-OpenBioLLM (70B) (Pal and Sankarasubbu, 2024), MedLlama3 (8B) (John Snow Labs, 2024), MedX (8B) (skumar9, 2024), HuatuoGPT-o1 (7B) (Chen et al., 2024), BioGPT (1.6B) (Luo et al., 2022), Llama-3.3-70B (Grattafiori et al., 2024) and GPT-4o-mini (Hurst et al., 2024), alongside three LightRAG-augmented variants that share the Llama-3.3-70B backbone and UniD³-KG but differ only in the LightRAG query mode (*local*, *global*, *mix*). This design isolates two factors of interest: the contribution of the biomedical knowledge graphs constructed by UniD³, and the choice of retrieval signal over those graphs.

4.3 Metrics

For DDM we report Coverage@ k , defined as $\text{Cov}@k = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}[\text{rel}(q) \cap \text{Top}_k(q) \neq \emptyset]$ for $k \in \{1, 2, 3\}$, alongside precision, recall and F1 against curated drug indications. For DEA we report accuracy, F1 and the macro-averaged area under the receiver-operating-characteristic curve (AUROC) across the four effectiveness levels, using the

Released artefact	Scale
PubMed articles processed	157,849
Task-specific source corpora	10,808 (DDM/DEA/DTA)
<i>Knowledge graphs (six total)</i>	
Stage I + Stage II per task	3 tasks $\times$ 2 stages
Total nodes (Stage II)	$\sim 112,700$
Total edges (Stage II)	$\sim 303,600$
<i>Open datasets</i>	
DDM pairs	28,915
DEA instances	15,042
DTA QA pairs	> 4,000
<i>Benchmark</i>	
No-retrieval LLM baselines	7
LightRAG-augmented variants	3
External validation suites	4 (CDR / DR / P3ps / MQA)
<i>Deployed system</i>	
UniD ³ -QA chatbot	public web interface

Table 1: Snapshot of artefacts released with this paper. All graphs, datasets, scripts and the chatbot are released at <https://github.com/QSong-github/UniD3>.

LLM-produced per-class scores for the AUROC computation. For DTA we follow the biomedical QA literature and report semantic similarity (SS, the cosine similarity between sentence-embedding representations of the predicted and reference answers), keyword match (KWM, the token-level F1 over biomedical keywords extracted from the predicted and reference answers), BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020); SS and KWM together cover lexical overlap and embedding-level faithfulness to the reference rationale.

5 Results

Table 1 summarises the artefacts released with this paper; the scores reported below are calibration points for downstream use rather than claims of method superiority.

5.1 Knowledge graphs constructed by UniD³

Running the pipeline of Section 3 produces six knowledge graphs (three tasks $\times$ two stages). Stage I aggregates over 10,808 PubMed articles and yields large but locally fragmented graphs dominated by “Other” biomedical entities (pathways, biomarkers and processes). Stage II re-extraction re-balances the type distribution towards the task-relevant axes (*drug*, *disease*, *target*) and substantially expands the relational structure. Edges grow from 72.0K to 126.0K on DDM (+75.0%), from 63.4K to 84.1K on DEA (+32.7%) and from

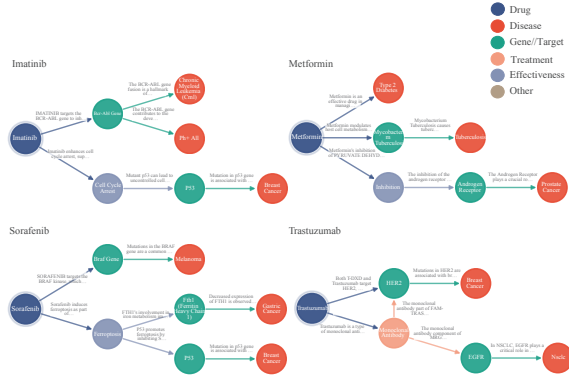


Figure 2: Multi-hop relation paths in the Stage II DDM KG for Imatinib, Trastuzumab, Sorafenib and Metformin, recovering known mechanism–indication chains and surfacing Metformin–tuberculosis as a repurposing candidate. An ego-network view of Cisplatin is given in Appendix D (Figure 4).

74.0K to 93.5K on DTA (+26.4%); the number of unique entities $|\mathcal{V}^{(t)}|$ contracts by 29–48% thanks to the fuzzy entity-merge step (Section 3.4), while graph density $\rho(\mathcal{G}^{(t)})$ rises by roughly an order of magnitude ($3.0\text{--}3.9 \times 10^{-5} \rightarrow 1.2\text{--}1.8 \times 10^{-4}$; Table 2; the detailed entity-type and relation breakdown is given in Figure 3 of Appendix D). The post-Stage II density remains well within the regime that LightRAG’s local and global retrievers index in real time, so the dual-stage gain does not trade off against retrieval-time tractability.

5.2 Drug-centric knowledge exploration

To probe the biomedical utility of $\mathcal{G}^{(\text{DDM})}$, we trace multi-hop relation paths around five widely studied agents (Figure 2; an extended Cisplatin ego-network is given in Appendix D, Figure 4). The Stage II DDM graph recovers canonical mechanism–indication chains for Imatinib (BCR-ABL $\rightarrow$ chronic myeloid leukaemia) (Druker et al., 2006), Trastuzumab (HER2 $\rightarrow$ breast cancer) (Slamon et al., 2001), Sorafenib (VEGFR/RAF $\rightarrow$ hepatocellular carcinoma) (Llovet et al., 2008) and Metformin (AMPK $\rightarrow$ tuberculosis as a repurposing candidate) (Singhal et al., 2014), alongside the DNA-damage–ferroptosis–platinum-resistance neighbourhood of Cisplatin (Dasari and Tchounwou, 2014; Ghosh, 2019; Dixon et al., 2012).

5.3 External validation of UniD³-Data

Table 3 reports precision, recall and F1 of the UniD³-KG triples against four public biomedical benchmarks. Across CDR, DrugReviews and P3ps the UniD³-Data DDM and DEA splits reach

$F1 \in [0.85, 0.87]$ with precision and recall both above 0.83, showing that prompt-driven extraction re-discovers expert-curated drug–disease and drug–effectiveness relations across three independently built corpora. MedicationQA is more demanding because answers are patient-friendly free text; DTA still attains $F1 = 0.8214$ at the BERTScore answer-matching level, within 4–5 points of the structured benchmarks. The three tasks are therefore *not* validated to the same standard: DDM and DEA are checked at the triple level against curated relations, whereas MedicationQA checks free-form DTA answers only, an indirect signal that does not establish whether a predicted drug–target relation is biologically correct. Triple-level DTA validation is outside this release’s scope.

Clinician verification. Three practising clinicians independently relabelled all 2,652 drug–indication pairs of a 469-drug FDA-approved, US-marketed sample as *effective* or *ineffective*, in randomised order and blind to the evidence the pipeline had selected. Inter-rater agreement is almost-perfect (Fleiss’ $\kappa = 0.954$) and majority-vote accuracy against UniD³-Data is 0.900; the protocol, per-rater statistics and the *ineffective*-class negative control are in Appendix E.

Robustness of the external agreement. Two further checks are reported in Appendix G. A corpus-overlap audit finds only 2 of CDR’s 1,499 source articles (0.13%) in our construction corpus, so the agreement is out-of-sample rather than rediscovery; and DEA disease–condition agreement moves by less than 0.5 pp across four effective/ineffective thresholds, with two licensed clinicians on a 46-item borderline sample agreeing at Cohen’s $\kappa = 0.88$ and matching the UniD³-Data label in 87% of decidable cases.

5.4 Benchmarking biomedical LLMs

Table 4 is a reference benchmark on the released UniD³-Data, comparing seven no-retrieval biomedical and general LLM baselines with three LightRAG-augmented Llama-3.3-70B variants that share UniD³-KG and differ only in the LightRAG query mode (*local*, *global*, *mix*).

Among the no-retrieval baselines, the strongest biomedical model is Llama3-OpenBioLLM (70B) on the structured tasks (Cov@1 = 0.7861, DEA F1 = 0.8305), while the strongest free-form rationale generator is Llama-3.3-70B on DTA (BERTScore = 0.9092). Smaller

Dataset	Research	Stage I					Stage II				
		Chunk	Token	Node	Edge	KG density	Chunk	Token	Node	Edge	KG density
DDM	3.6K	5.8K	30.4M	64.3K	72.0K	3.48×10^{-5}	16.7K	3.4M	45.9K	126.0K	1.20×10^{-4}
DEA	3.5K	4.7K	22.1M	56.9K	63.4K	3.91×10^{-5}	10.8K	2.3M	30.5K	84.1K	1.81×10^{-4}
DTA	3.7K	5.8K	29.6M	69.9K	74.0K	3.03×10^{-5}	13.0K	2.2M	36.3K	93.5K	1.42×10^{-4}

Table 2: Statistics in the dual-stage entity extraction process across the DDM, DEA and DTA knowledge graphs. Stage I performs chunk-level extraction at the paper scale; Stage II re-extracts over each entity’s KG-level neighbourhood. Stage II expands edges by $\sim 1.3\text{--}1.7\times$ while raising graph density by roughly an order of magnitude, recovering cross-document evidence missed in Stage I.

Benchmark	DDM			DEA			DTA		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
CDR	0.8302	0.8778	0.8526	0.8495	0.8851	0.8668	–	–	–
DrugReviews	0.8553	0.8923	0.8727	0.8479	0.8885	0.8677	–	–	–
P3ps	0.8486	0.8951	0.8707	0.8494	0.8920	0.8696	–	–	–
MedicationQA	–	–	–	–	–	–	0.8147	0.8287	0.8214

Table 3: External validation of UniD³ datasets against four public biomedical benchmarks. P/R/F1 are triple-level for DDM and DEA, and answer-level (BERTScore matching) for DTA on MedicationQA. DEA labels on CDR/DrugReviews/P3ps project indication-level associations to a binary effective/ineffective scheme (Section 4.1); blank cells mean the benchmark lacks that task’s pairs. F1 stays ≥ 0.85 on the three indication-style benchmarks and 0.82 on free-form drug QA.

medical-specialised models (BioGPT, HuatuoGPT-o1, MedX, MedLlama3) lag substantially, especially on Cov@2 / Cov@3 and on DTA keyword matching, indicating that domain specialisation alone is insufficient when the task requires multi-hop biomedical reasoning over relations not co-located in a single sentence.

Adding retrieval over UniD³-KG improves every metric on every task under the global and Mix variants, and improves nearly every metric under the local-only variant (the only exception being a marginal 0.69 pp decrease in DTA BLEU). The local-only variant already lifts DDM Cov@3 from 0.4311 to 0.7655 and DTA KWM from 0.4695 to 0.6715 over the Llama-3.3-70B base, showing that the constructed KGs carry biomedical signal that the underlying LLM does not encode even at 70B scale.

The *Mix* retriever, LightRAG’s built-in mode that integrates entity-anchored, community-level and chunk-level evidence in a single retrieval call, serves as the upper-bound reference variant within the released benchmark and scores highest on every metric. Of its gains over the strongest no-retrieval baseline (Llama-3.3-70B), only the +40.8 pp on DDM Cov@3 is externally grounded against curated indications; the corresponding +21.2 pp on DEA F1 and +38.0 pp on DTA KWM are self-consistency upper bounds (Limitations), with independent counterparts in Table 3.

5.5 Does the knowledge graph beat text-only retrieval?

Table 4 contrasts no-retrieval LLMs with UniD³-KG-backed retrieval, which shows that retrieval helps but not that the *graph* is what helps. Holding the corpus, the Llama-3.3-70B backbone and the scorer fixed and varying only how chunks are ranked, the dual-stage KG leads lexical Okapi BM25 by 7.1 to 7.7 pp on Cov@1/2/3 and dense LightRAG-*naive* retrieval by 11.6 to 14.3 pp, paired on the full 1,710-drug DDM set, with bootstrap CIs excluding zero and McNemar $p < 0.001$ throughout (Appendix G, Table 5). BM25 itself beats dense retrieval, so this is not a straw man. Comparison against curated biomedical KGs such as Hetionet or ROBOKOP is a separate cross-resource question we do not address here.

5.6 Ablation: dual-stage vs. single-stage extraction

Table 6 (Appendix G) isolates the contribution of Stage II re-extraction by re-running the full benchmark with *Stage I-only* graphs (i.e. omitting the KG-level pass of Stage II). Removing Stage II degrades every Mix-retriever metric. DDM Cov@3 drops by 14.2 points (0.8388 $\rightarrow$ 0.6968), DEA F1 by 11.4 points and DTA BERTScore by 3.1 points, while precision on the external CDR, DrugReviews and P3ps benchmarks decreases by 3–6 points. The drop is largest on the multi-hop Cov@3 metric and

Dataset	DDM				DEA [†]			DTA [†]			
Metrics	Cov@1	Cov@2	Cov@3	ACC	F1	AUROC	SS	KWM	BLEU	ROUGE	BERTsc
Llama3-OpenBioLLM (70B)	0.7861	0.7015	0.5133	0.8277	0.8305	0.8657	0.8428	0.3811	0.4343	0.3977	0.8591
BioGPT (1.6B)	0.4211	0.2879	0.1613	0.4964	0.6072	0.5311	0.5183	0.2275	0.2345	0.2079	0.8241
HuatuoGPT-o1 (7B)	0.4577	0.3195	0.2006	0.5598	0.6726	0.5883	0.2838	0.2407	0.2410	0.2361	0.8273
MedX (8B)	0.6827	0.4551	0.3635	0.6544	0.5963	0.6514	0.3667	0.2073	0.2337	0.2894	0.8169
MedLlama3 (8B)	0.5177	0.3872	0.2119	0.5753	0.6058	0.5778	0.3101	0.1818	0.2461	0.2511	0.8378
GPT-4o-mini (NA)	0.7522	0.6332	0.5001	0.7930	0.7568	0.7895	0.8112	0.3962	0.4223	0.4490	0.8682
Llama-3.3 (70B)	0.7981	0.6298	0.4311	0.7641	0.7762	0.7662	0.8834	0.4695	0.5435	0.4315	0.9092
Llama-3.3-LightRAG-local	0.8437	0.7958	0.7655	0.8491	0.8627	0.8433	0.8957	0.6715	0.5366	0.6385	0.9418
Llama-3.3-LightRAG-global	0.8967	0.8325	0.7968	0.8635	0.8589	0.8372	0.9369	0.6997	0.6218	0.6739	0.9677
Llama-3.3-LightRAG-mix	0.9437	0.8705	0.8388	0.9764	0.9883	0.9899	0.9811	0.8497	0.7815	0.7568	0.9871

Table 4: Reference benchmark on the released UniD³ datasets, *not* a method-superiority claim. [†]The DEA and DTA columns are self-consistency upper bounds because candidates and references both come from the same KG-driven pipeline; only DDM Cov@*k* is independently grounded against curated indications (Table 3). The three LightRAG variants share the Llama-3.3-70B backbone and the UniD³ graphs, differing only in query mode, so *Mix* is the within-benchmark upper-bound reference. Cov@*k* is coverage between the real and predicted disease/effectiveness label; SS = semantic similarity, KWM = keyword matching, ROUGE = ROUGE-L, BERTsc = BERTScore. Best per metric in bold.

on KWM-style answer faithfulness, which matches our hypothesis from Section 3.4: cross-chunk and cross-document edges are precisely what the second pass recovers, and these edges disproportionately support multi-hop reasoning over biomedical mechanisms.

We additionally ablate the *prompt agreement filter* (Section 3.3). Replacing the four-LLM-agreement filter with the raw Llama-3.3-70B exemplars reduces DEA F1 by 5.7 points and increases the rate of malformed JSON triples by $\approx 9\%$, indicating that the cross-LLM agreement step is a cheap but effective regulariser against single-model prompt artefacts.

5.7 Error analysis

We sampled 300 DDM and 200 DTA predictions of the Mix variant that disagree with the curated benchmarks and categorised them by hand (Appendix H). Synonym splits ($\sim 28\%$) and long-tail entities ($\sim 22\%$) dominate the residual errors, alongside indication-vs-contraindication confusion ($\sim 19\%$), label granularity mismatch ($\sim 17\%$) and plausible repurposing candidates ($\sim 14\%$, e.g. metformin–tuberculosis (Singhal et al., 2014)) that literature supports but the benchmarks omit. The latter three reflect benchmark limitations rather than model failure, so the F1 in Table 3 is a conservative lower bound on true biomedical correctness.

Because synonym splits are the single largest residual category, normalised Levenshtein matching (§3.4) is not by itself an adequate identity criterion for biomedical entities. Linking a 500-entity sample per type to canonical public identifiers an-

chors 46.4% of drugs to RxNorm, 56.4% of diseases to MeSH and 78.4% of genes to NCBI Gene, so we adopt ontology identifiers as the canonical merge key on top of the Levenshtein step, a post-hoc re-keying that needs no GPU time or re-extraction (Appendix G).

5.8 Efficiency and scalability

End-to-end construction of the six graphs and three datasets from 157,849 articles consumed $\sim 3,530$ A100-hours on a single A100-80G (59% of which is Stage II re-extraction; per-task breakdowns in Appendix J). That is a one-time construction cost, not a cost of *using* the release: consumers query the published graphs and datasets directly at 3.6 ± 0.4 s per UniD³-QA query on one A100 (1.5 s of it graph retrieval), and never rebuild. Updates are incremental, since Stage I processes only new articles and Stage II continues from those outputs, so cost scales with new literature rather than corpus size.

5.9 Case study: UniD³-QA chatbot

We deploy the framework as UniD³-QA, an explainable chatbot pairing every answer with the supporting Stage II KG path π and PubMed evidence sentences. On 50 PubMedQA-derived questions, three independent biomedical reviewers preferred UniD³-QA over the base Llama-3.3-70B in 78% of cases, citing better grounding and traceable mechanism chains. The small sample and absence of formal inter-rater statistics make this indicative rather than confirmatory (transcripts in Appendix O).

6 Discussion

The dual-stage strategy is the largest source of structural improvement in our experiments. Stage I is competitive on chunk-local relations, but Stage II brings cross-document mechanism paths (e.g. ferroptosis-mediated platinum effects; Section 5.2) and recovers FDA-aligned indications that no single PubMed article states explicitly.

Overlap with curated repositories such as DrugBank is here a desirable property rather than test-set leakage: the 0.90 effectiveness agreement on FDA-approved drugs (Section 5.3) confirms in-distribution reliability, while literature-supported examples (metformin–tuberculosis, cisplatin–ferroptosis; Section 5.2) reframe the remainder as candidate repurposing hypotheses for expert review.

Compared with curated/supervised pipelines (Lai et al., 2023; Davis et al., 2009), UniD³ trades per-relation verifiability for breadth and update speed (further trade-offs in Limitations). The Mix retriever’s externally-grounded gain on DDM Cov@3 (Table 3) is largest where the LLM is weakest, suggesting that a high-quality task-specific KG yields a larger return than scaling up a generic biomedical LLM.

7 Conclusion

UniD³ releases UniD³-KG (six knowledge graphs), UniD³-Data (three QA datasets with KG paths) and UniD³-QA (an explainable chatbot), distilled from 157,849 PubMed articles. Retrieval over UniD³-KG consistently improves DDM/DEA/DTA on external benchmarks, with Stage II the largest single ablation gap (Table 6); a three-clinician verification on 469 FDA-approved drugs (majority-vote accuracy 0.90, Fleiss’ $\kappa = 0.95$) confirms reliability for biomedical-NLP and drug-repurposing work.

Limitations

UniD³ has several limitations that bound the claims we make in this paper. The corpus consists exclusively of English-language PMC Open Access full-text articles; clinical notes, non-PMC biomedical repositories and non-English literature are not covered, and recovering relations from those sources will likely require additional preprocessing and prompt adaptation. All extraction is driven by a single backbone (Llama-3.3-70B), and LLM-based curation is known to inherit factual and stylistic biases from the underlying model; neither the schema

nor the few-shot exemplars can fully neutralise hallucinations in long-tail biomedical entities, and complementary claim-level verification (Ji et al., 2026) would be a natural addition for downstream deployments. The benchmark’s value lies in comparing LLM baselines against the released datasets rather than in claiming a specific retrieval variant is the best method. The Mix-retriever rows in Table 4 reach near-ceiling scores on DEA F1 and DTA BERTScore because their candidates and the reference answers come from the same KG-driven pipeline, a self-consistency upper bound rather than independent ground truth. The cross-LLM ranking on the same datasets, the external-benchmark figures of Table 3 and the clinician verification on FDA-approved drugs (Section 5.3) provide the independent evidence of dataset quality. Entity merging relies on normalised Levenshtein similarity rather than canonical ontology IDs (MeSH, UMLS, RxNorm, DrugBank); this catches the majority of surface variants but cannot resolve acronym–expansion pairs, salt-form variants or hierarchical label mismatches. We report a partial ontology grounding (§5.7) and adopt canonical identifiers as the merge key, but the drug and disease anchor rates of 46% and 56% mean a substantial long tail remains unlinked, and richer benchmarking against curated biomedical KGs such as Hetionet or ROBOKOP is left to future work. The clinician review focuses on the FDA-approved, US-marketed subset of the DDM dataset; coverage of off-label uses, non-US-marketed drugs and the DEA/DTA datasets remains a prerequisite for any clinical decision-support deployment that builds on UniD³. Despite fuzzy entity merging, the constructed KGs still contain residual synonym splits and missing edges, particularly for emerging drug classes (Appendix K). Running the full pipeline over 157,849 articles is also computationally non-trivial; we report total GPU hours and dollar-cost estimates in Appendix J so that practitioners can plan replication or extensions accordingly. For groups that need a cheaper extractor rather than the released artefacts, distilling the backbone into a compact student is the established route, and has been shown to preserve clinical utility under the fixed on-premises budgets typical of hospital deployments (Lan et al., 2026a); analogous self-distillation methods are now being developed for agentic pipelines (Lan et al., 2026b). Finally, UniD³ is a text-only resource: it represents what the literature states, not imaging, motion or other non-textual clinical sig-

nals. Multimodal models that reason over such signals, for example over paediatric gait recordings (Shen et al., 2026) or in broader multimodal sensing (Li et al., 2026; Yu et al., 2025), and methods for judging whether generated content is authentic (Fu et al., 2025), address evidence types our graphs do not cover, and integrating them is outside this release.

Validation coverage. The clinician study covers the DDM split on FDA-approved, US-marketed drugs. Off-label indications and drugs not marketed in the US are outside the current release, as is a dedicated clinician-graded DEA/DTA gold set; DEA and DTA rest on external benchmark agreement and, for DTA, on answer-level matching only.

Ethical Considerations

All source text used by UniD³ comes from publicly available PubMed articles; no patient-level, identifiable or otherwise sensitive clinical data is involved at any stage of pipeline construction or evaluation. The released datasets, knowledge graphs and chatbot are intended strictly for biomedical research and do not constitute medical advice, diagnosis or treatment recommendation. Because the framework can surface candidate drug–disease associations not present in curated FDA and DrugBank lists, it has dual-use potential for unsupervised drug repurposing. We mitigate this risk through three complementary measures: every generated answer is grounded in retrieved KG paths and source sentences so that biomedical experts can audit each suggestion; the chatbot interface emphasises that all outputs are research aids rather than clinical guidance; and we recommend that downstream users maintain a human-in-the-loop review process before acting on any UniD³-derived hypothesis. The annotation work performed by the participating clinician was conducted on a voluntary basis under the institutional guidelines of their home institution.

Use of AI assistants. The authors used AI assistants (Claude, GitHub Copilot) for code drafting and manuscript language polishing. All scientific claims, citations, dataset constructions and experimental results were reviewed and verified by the authors, who take full responsibility for the final content.

Acknowledgments

Q.S. is supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM151089. This work was also supported by NSF ACCESS “Discover” Allocation Award No. CIS261195.

References

- Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers’ medication questions and trusted answers. *MedInfo*, 264:25–29.
- Yelaman Abdullin, Diego Molla, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2023. [Synthetic dialogue dataset generation using LLM agents](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 181–191. Association for Computational Linguistics.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.
- Donald C Comeau, Chih-Hsuan Wei, Rezarta Islamaj Doğan, and Zhiyong Lu. 2019. [PMC text mining subset in BioC: about three million full-text articles and growing](#). *Bioinformatics*, 35(18):3533–3535.
- Jacqueline Corrigan-Curay, Leonard Sacks, and Janet Woodcock. 2018. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *Jama*, 320(9):867–868.
- Udptaman Das, Krishnasai B Atmakuri, Duy Ho, Chi Lee, and Yugyung Lee. 2026. Clinical knowledge graph construction and evaluation with multi-llms via retrieval-augmented generation. *arXiv preprint arXiv:2601.01844*.
- Shaloam Dasari and Paul Bernard Tchounwou. 2014. Cisplatin in cancer therapy: molecular mechanisms of action. *European journal of pharmacology*, 740:364–378.
- Allan Peter Davis, Cynthia G Murphy, Cynthia A Saraceni-Richards, Michael C Rosenstein, Thomas C Wiegers, and Carolyn J Mattingly. 2009. Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic acids research*, 37(suppl_1):D786–D792.
- Scott J Dixon, Kathryn M Lemberg, Michael R Lamprecht, Rachid Skouta, Eleina M Zaitsev, Caroline E Gleason, Darpan N Patel, Andras J Bauer, Alexandra M Cantley, Wan Seok Yang, and 1 others. 2012. Ferroptosis: an iron-dependent form of nonapoptotic cell death. *cell*, 149(5):1060–1072.

- Brian J Druker, François Guilhot, Stephen G O'Brien, Insa Gathmann, Hagop Kantarjian, Norbert Gattermann, Michael WN Deininger, Richard T Silver, John M Goldman, Richard M Stone, and 1 others. 2006. Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *New England Journal of Medicine*, 355(23):2408–2417.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.
- Tianyu Fan, Jingyuan Wang, Xubin Ren, and Chao Huang. 2025. Minirag: Towards extremely simple retrieval-augmented generation. *arXiv preprint arXiv:2501.06713*.
- Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, and 1 others. 2025. Learning human-perceived fakeness in ai-generated videos via multimodal llms. *arXiv preprint arXiv:2509.22646*.
- Sumit Ghosh. 2019. Cisplatin: The first metal based anticancer drug. *Bioorganic chemistry*, 88:102925.
- Felix Gräber, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In *Proceedings of the 2018 international conference on digital health*, pages 121–125.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. **LightRAG: Simple and fast retrieval-augmented generation**. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10746–10761. Association for Computational Linguistics.
- Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A Rossi, Subhabrata Mukherjee, Xianfeng Tang, and 1 others. 2024. Retrieval-augmented generation with graphs (graphrag). *arXiv preprint arXiv:2501.00309*.
- Daniel Scott Himmelstein, Antoine Lizée, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *elife*, 6:e26726.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Yuelyu Ji, Min Gu Kwak, Hang Zhang, Xizhi Wu, Chenyu Li, and Yanshan Wang. 2026. Medragchecker: Claim-level verification for biomedical retrieval-augmented generation. *arXiv preprint arXiv:2601.06519*.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. **PubMedQA: A dataset for biomedical research question answering**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577. Association for Computational Linguistics.
- John Snow Labs. 2024. JSL-MedLlama-3-8B v2.0. Hugging Face model card. <https://huggingface.co/johnsnowlabs/JSL-MedLlama-3-8B-v2.0>.
- Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Qingyu Chen, and Zhiyong Lu. 2023. Biorex: improving biomedical relation extraction by leveraging heterogeneous datasets. *Journal of Biomedical Informatics*, 146:104487.
- Qizhen Lan, Aaron Choi, Jun Ma, Bo Wang, Zhongming Zhao, Xiaoqian Jiang, and Yu-Chun Hsu. 2026a. From performance to practice: Knowledge-distilled segmentator for on-premises clinical workflows. *npj Health Systems*.
- Qizhen Lan, Xi Xiao, Xiangchen Guan, Mengchen Fan, Moule Lin, Jung Im Choi, and Lijing Zhu. 2026b. Trust is not enough: Influence calibration for on-policy self-distillation in agentic RL. *arXiv preprint arXiv:2608.14945*.
- Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhao Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, and 1 others. 2026. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*.
- Dawei Li, Shu Yang, Zhen Tan, Jae Young Baik, Sukwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding, and 1 others. 2024. **DALK: Dynamic co-augmentation of LLMs and KG to answer Alzheimer's disease questions with scientific literature**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2187–2205. Association for Computational Linguistics.
- Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreave v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

- Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. 2025. Biomedrag: A retrieval augmented large language model for biomedicine. *Journal of Biomedical Informatics*, 162:104769.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.
- Xiaona Liu, Qing Wang, Minghao Zhou, Yanfei Wang, Xuefeng Wang, Xiaobo Zhou, and Qianqian Song. 2024. Drugformer: Graph-enhanced language model to predict drug sensitivity. *Advanced Science*, 11(40):2405861.
- Josep M Llovet, Sergio Ricci, Vincenzo Mazzaferro, Philip Hilgard, Edward Gane, Jean-Frédéric Blanc, Andre Cosme De Oliveira, Armando Santoro, Jean-Luc Raoul, Alejandro Forner, and 1 others. 2008. Sorafenib in advanced hepatocellular carcinoma. *New England journal of medicine*, 359(4):378–390.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.
- Nicholas Matsumoto, Jay Moran, Hyunjun Choi, Miguel E Hernandez, Mythreye Venkatesan, Paul Wang, and Jason H Moore. 2024. Kragen: a knowledge graph-enhanced rag framework for biomedical problem solving using large language models. *Bioinformatics*, 40(6):btac353.
- Guofeng Meng, Li Shen, Qiuyan Zhong, Wei Wang, Haizhou Zhang, and Xiaozhen Wang. 2025. fastbmrag: A fast graph-based rag framework for efficient processing of large-scale biomedical literature. *arXiv preprint arXiv:2511.10014*.
- Kenneth Morton, Patrick Wang, Chris Bizon, Steven Cox, James Balhoff, Yaphet Kebede, Karamarie Fecho, and Alexander Tropsha. 2019. Robokop: an abstraction layer and user interface for knowledge graphs to support question answering. *Bioinformatics*, 35(24):5382–5384.
- David Ochoa, Andrew Hercules, Miguel Carmona, Daniel Suveges, Asier Gonzalez-Uriarte, Cinzia Malangone, Alfredo Miranda, Luca Fumis, Denise Carvalho-Silva, Michaela Spitzer, and 1 others. 2021. Open targets platform: supporting systematic drug-target identification and prioritisation. *Nucleic acids research*, 49(D1):D1302–D1310.
- P3ps. 2024. Condition-to-drug mapping dataset. Hugging Face dataset. https://huggingface.co/datasets/P3ps/condition_to_drug.
- Ankit Pal and Malaikannan Sankarasubbu. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. <https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B>.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.
- Nabeel Seedat, Nicolas Huynh, Boris Van Breugel, and Mihaela Van Der Schaar. 2024. **Curated LLM: Synergy of LLMs and data curation for tabular augmentation in low-data regimes**. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 44060–44092. PMLR.
- Yifan Shen, Boyi Li, Meihuan Huang, Yuanzhe Liu, Xu Cao, Jinyang Jin, Zhengyuan Li, Anglin Liu, Junho Kim, Jingyuan Zhu, and 1 others. 2026. Decoding children’s gait behavior. *arXiv preprint arXiv:2608.00371*.
- Jake Silberg, Kyle Swanson, Elana Simon, Angela Zhang, Zaniar Ghazizadeh, Scott Ogden, Hisham Hamadeh, and James Zou. 2024. Unitox: leveraging llms to curate a unified dataset of drug-induced toxicity from fda labels. *Advances in Neural Information Processing Systems*, 37:12078–12093.
- Amit Singhal, Liu Jie, Pavanish Kumar, Gan Suay Hong, Melvin Khee-Shing Leow, Bhairav Paleja, Liana Tsenova, Natalia Kurepina, Jinmiao Chen, Francesca Zolezzi, and 1 others. 2014. Metformin as adjunct antituberculosis therapy. *Science translational medicine*, 6(263):263ra159–263ra159.
- skumar9. 2024. Llama-medx_v3.2. Hugging Face model repository, fine-tuned from Meta-Llama-3.1-8B. https://huggingface.co/skumar9/Llama-medx_v3.2, accessed 21 August 2026.
- Dennis J Slamon, Brian Leyland-Jones, Steven Shak, Hank Fuchs, Virginia Paton, Alex Bajamonde, Thomas Fleming, Wolfgang Eiermann, Janet Wolter, Mark Pegram, and 1 others. 2001. Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that overexpresses her2. *New England journal of medicine*, 344(11):783–792.
- ZiaurRehman Tanoli, Zaid Alam, Markus Vähä-Koskela, Balaguru Ravikumar, Alina Malyutina, Alok Jaiswal, Jing Tang, Krister Wennerberg, and Tero Aittokallio. 2018. Drug target commons 2.0: a community platform for systematic analysis of drug-target interaction profiles. *Database*, 2018:bay083.
- Qing Wang, Bo Li, Jialu Liang, Daling Shi, Bob Zhang, and Qianqian Song. 2026. DrugClaw and DrugAudit: A primary-source-grounded agent and authority-aware benchmark for drug-information question answering. *arXiv preprint arXiv:2606.01434*.
- Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. **A**

[prompt pattern catalog to enhance prompt engineering with ChatGPT](#). In *Proceedings of the 30th Conference on Pattern Languages of Programs (PLOP)*. The Hillside Group.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, and 1 others. 2018. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. [Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467. Association for Computational Linguistics.

Jihong Yang, Zheng Li, Xiaohui Fan, and Yiyu Cheng. 2014. Drug–disease association and drug-repositioning predictions in complex diseases using causal inference–probabilistic matrix factorization. *Journal of chemical information and modeling*, 54(9):2562–2569.

Muhammed A Yildirim, Kwang-Il Goh, Michael E Cusick, Albert-Laszlo Barabasi, and Marc Vidal. 2007. Drug-target network. *Nature biotechnology*, 25(10):1119–1126.

Tianjiao Yu, Xinzhuo Li, Yifan Shen, Yuanzhe Liu, and Ismini Lourentzou. 2025. Core3d: Collaborative reasoning as a foundation for 3d intelligence. *arXiv preprint arXiv:2512.12768*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations (ICLR)*.

Wen Zhang, Xiang Yue, Weiran Lin, Wenjian Wu, Ruoqi Liu, Feng Huang, and Feng Liu. 2018. Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC bioinformatics*, 19(1):233.

Yongjun Zhu, Olivier Elemento, Jyotishman Pathak, and Fei Wang. 2019. Drug knowledge bases and their applications in biomedical informatics research. *Briefings in bioinformatics*, 20(4):1308–1321.

A Prompt Templates

This appendix reproduces the three families of prompts used by UniD³: the task-specific classification prompt $\mathcal{P}_{\text{cls}}$ of Section 3.2, the entity and relation extraction prompts $\mathcal{P}_t$ of Section 3.3, and the dataset-generation prompts $\mathcal{P}_t^{\text{gen}}$ of Section 3.5. Templates are reproduced verbatim from the released software bundle except for one mitigation noted below.

Classification prompt $\mathcal{P}_{\text{cls}}$. LLM _{θ} is invoked on each retrieved PubMed title under the following template, which routes the title to one of the three tasks or to an “irrelevant” bucket.

You should think as an expert in the medical field. I will give you a bunch of research paper titles. Please select the appropriate papers for the three tasks I set. You need to classify each paper title. If you think a paper does not belong to any task, you can classify it as “irrelevant”. The three tasks are:

Task 1 (Drug Disease Matching). Identify potential drugs that could treat [specific disease]. For each drug, provide an analysis of its target genes and pathways and explain their relevance to the disease mechanism.

Task 2 (Drug Effectiveness Assessment). Evaluate the effectiveness of [specific drug] for treating [specific disease]. Include an analysis of clinical or preclinical data, the strength of the drug’s interaction with its target genes or pathways, and any evidence of therapeutic outcomes.

Task 3 (Drug Target Analysis). Map the genes and pathways targeted by [specific drug]. Explain how these targets are implicated in [specific disease] and assess whether the drug’s mechanism of action aligns with the disease’s underlying biology.

Output format: Drug Disease
Matching: [title1, title2, ...];
Drug Effectiveness
Assessment: [...]; Drug Target
Analysis: [...].

Extraction prompt $\mathcal{P}_t$. For each task t , LLM _{θ} extracts typed triples from each chunk under the template below, instantiated with the task-specific instruction $\mathcal{I}_t$ (the task name and definition from $\mathcal{P}_{\text{cls}}$ above). The submitted template tones down one sentence in the original prompt. The original requested “the maximum computing power and the ultimate depth of analysis, ... show your true cognitive limit”, which we replace by a neutral request for substantive rather than surface analysis; the rest of the template is verbatim.

In a task named $\mathcal{I}_t$, you are an expert in drugs and diseases. Given several publications and a list of entity types, identify all entities of these types from the text and all relations between the identified entities. Focus only on relations with the listed entities. Provide a thorough analysis grounded in the cited publications, identifying substantive findings rather than surface descriptions. Please think step by step to ensure correctness.

Few-shot agreement filter. For every task t , three biomedical-NLP annotators independently wrote candidate few-shot exemplars $\mathcal{F}_t^{\text{seed}}$. Each exemplar was then prompted to four LLMs (LLM _{θ} (Llama-3.3-70B), GPT-5.4, Claude-4.6-Sonnet and

Gemini-3-Pro). Only exemplars whose extracted triples agreed across all four models (after the same fuzzy entity-merge step described in Section 3.4) were retained, yielding the final exemplar pool $\mathcal{F}_t$ of 9–12 examples per task; the retained exemplars in the LightRAG triple format, together with the source biomedical input texts and the per-LLM extraction outputs, are released with the candidate-output bundle. This filter trades some coverage for robustness against prompt-induced bias from any single model.

Hierarchical keyword extraction. At query time the LightRAG retriever decomposes a natural-language question into high-level keywords (broad concepts: drug name, disease name, analytical objective) and low-level keywords (specific molecular targets, biological mechanisms, clinical endpoints). The decomposition covers diverse query types in the released benchmark, including condition-anchored queries (“potential drugs for Alzheimer’s”), drug-anchored queries (“effectiveness of Remdesivir for COVID-19”) and target-pathway queries (“target pathways for Trastuzumab in HER2-positive breast cancer”); the per-query high-level/low-level keyword bundles are released with the candidate-output bundle.

Dataset-generation prompt $\mathcal{P}_t^{\text{gen}}$. The three task-specific dataset-generation prompts share a common header (task instruction + worked OMADACYCLINE example) and differ only in the output schema. The DDM instance is reproduced below.

In a task named Drug-Disease Matching, identify potential diseases that could be treated by a [specific Drug]. Given the drug name, respond with diseases the drug can treat or relieve and provide a brief explanation; for each drug, analyse its target genes and pathways and explain their relevance to the disease mechanism. Format the response as a drug summary followed by a list of diseases (with proven efficacy and under investigation) and their explanations.

Worked example. Drug name: OMADACYCLINE. *Drug summary:* Omadacycline is a novel aminomethylcycline antibiotic used to treat several bacterial infections; its primary mechanism is binding to the 30S ribosomal subunit, inhibiting bacterial protein synthesis and preventing bacterial growth. *Diseases:* Community-Acquired Bacterial Pneumonia (CABP), because omadacycline inhibits protein synthesis in *Streptococcus pneumoniae* and limits proliferation in the lungs; Acute Bacterial Skin and Skin Structure Infections (ABSSI), because it targets *Staphylococcus aureus* including MRSA; and Urinary Tract Infections (under investigation), because it sup-

presses *Escherichia coli* protein production in the urinary tract.

The DEA instance replaces the disease-list output with two parallel lists (“effectively treated” and “ineffective or unclear efficacy”) and a short explanation; the DTA instance replaces it with a single Question–Answer pair that maps molecular targets to indications and assesses biological plausibility. Both variants re-use the same OMADACYCLINE example structure as the DDM template; the verbatim text of all three is in the released candidate-output bundle.

B PubMed Retrieval and Preprocessing

We retrieve articles from the PubMed Central Open Access (PMC OA) Subset (Comeau et al., 2019), the freely redistributable subset of PubMed full-text indexed via NCBI’s E-utilities API; we use the keyword `Drug` and download all available records up to April 2025. Restricting to PMC OA ensures that the constructed knowledge graphs and the released evidence sentences carry redistribution rights compatible with public dataset release. Records are de-duplicated by PMID, filtered to those with full-text content available, and language-filtered to English. Articles shorter than 80 tokens (after basic whitespace tokenisation) are discarded as too information-poor for relation extraction, and articles longer than 2,048 tokens are truncated at the section boundary closest to the limit to fit within the LightRAG chunking budget (Appendix C). The resulting corpus contains $|\mathcal{D}| = 157,849$ articles. After task-specific classification (Section 3.2), the per-task training corpora $\mathcal{D}_{\text{src}}^{(t)}$ contain 3,608 DDM, 3,512 DEA and 3,688 DTA articles; articles assigned to “Other” are dropped from downstream extraction but kept as a public auxiliary set for future biomedical NLP work.

C KG Construction Hyper-parameters

We use the public LightRAG implementation (Guo et al., 2025) with the following settings: chunk size 1,024 tokens with a 128-token overlap; entity-extraction temperature 0.0 for determinism; maximum output tokens 2,048 per chunk; embedding model `bge-large-en-v1.5` for retrieval; community detection via Leiden with resolution 1.0. Stage II uses $h = 2$ -hop neighbourhoods serialised in canonical JSON order, capped at 4,096 tokens per KG-level chunk to keep the context window manageable. The fuzzy-matching

threshold for Merge $_{\tau}$ is set to $\tau = 0.92$ on the normalised Levenshtein similarity $\text{sim}(e, e') = 1 - \text{Lev}(e, e') / \max(|e|, |e'|)$ implemented through RapidFuzz; this value was selected on a held-out validation set of 200 manually-merged drug names. The Mix retriever uses LightRAG’s default mode="mix" setting; we do not modify its internal combination of entity, community and chunk signals. All extractions and benchmarks are run with Llama-3.3-70B on a single NVIDIA A100-80G GPU. We report seed-level variance across 3 runs in Appendix F.

D Full KG Statistics

Table 2 summarises the dual-stage statistics in the main body. We additionally release, with the dataset bundle, a machine-readable manifest that records, for each of the six UniD³ graphs, per-entity-type counts (drug, disease, target, biomarker, pathway, mechanism, side-effect, outcome, other), per-relation-type counts (treats, induces, modulates, interacts-with, and similar), the distribution of node degrees and shortest-path lengths, and the provenance of every triple in the form of source PMID, source sentence span and extraction stage. Entity-type composition is dominated by “Other” in Stage I ($\approx 58\%$) and re-balances toward task-relevant axes in Stage II (*drug*: 11% $\rightarrow$ 22%; *disease*: 14% $\rightarrow$ 25%; *target*: 9% $\rightarrow$ 17% on DTA), as visualised in Figure 3.

E Clinician Review Protocol

Sampling. We restricted the UniD³-Data DDM predictions to drugs that are FDA-approved and currently marketed in the United States, yielding 469 unique drugs after de-duplication. Despite minor spelling errors in a small subset of extracted names, most could be matched to official drug names through manual review.

Recruitment and instructions. Three practising clinicians were recruited from the authors’ clinical collaborator network at the participating institutions. One rated the set first; two further clinicians later re-rated every item independently, in randomised order and without being shown the pipeline-selected evidence, which makes their labels independent annotation rather than evidence-supported verification. All three received the following written instructions verbatim (the two added raters saw the same text with the evidence clause removed):

“For each of 469 FDA-approved, US-marketed drugs you will be shown the indication(s) predicted by an LLM-driven pipeline and the supporting PubMed evidence sentences. Independently of any pipeline label, mark each drug as EFFECTIVE or INEFFECTIVE for the listed indication based on FDA-approved labels, current clinical guidelines and your own expertise. No patient-level data is shown; this is a research-only task and the resulting annotations will be used solely to evaluate the pipeline’s drug-effectiveness predictions. There are no known risks to participation, which is voluntary and uncompensated. You may decline any item or withdraw at any time.”

Rubric. Each clinician assessed the 469 drugs, covering 2,652 drug–indication pairs in total, as *effective* or *ineffective* for the indication(s) predicted by UniD³, based on FDA-approved labels, clinical guidelines and their own expertise. All raters were blinded to the UniD³ label during scoring. Item order was randomised per rater, adjudication is by majority vote, and the adjudicated label was then compared against the corresponding UniD³-Data DDM split annotation.

Agreement. Inter-rater reliability over the three clinicians is Fleiss’ $\kappa = 0.954$ (95% CI [0.944, 0.963]), with pairwise Cohen’s κ of 0.960, 0.970 and 0.931; per-rater accuracy against UniD³-Data is 0.900, 0.885 and 0.885. Treating the adjudicated expert labels as ground truth and the UniD³-DDM dataset annotations as predictions, the verification yields majority-vote accuracy = 0.900 ([0.890, 0.911]), and for the first rater precision = 0.88, recall = 0.92, F1 = 0.90 and AUROC = 0.90. As controls, the *effective* class ($n = 1,303$) reaches majority accuracy 0.920 ($\kappa = 0.841$) and the *ineffective* class ($n = 1,349$) 0.880 ($\kappa = 0.910$). The strong agreement across all five metrics supports the framing in Section 5.3 that the UniD³-Data DDM split is robust and suitable for drug-effectiveness classification.

Compensation and ethics. The clinician’s participation was voluntary and uncompensated, conducted under the institutional research guidelines of the clinician’s home institution. No patient-level data was shown at any point of the review; all evidence sentences were drawn from public PubMed articles.

F Full Benchmark Tables

We release a machine-readable benchmark bundle that, for every cell of Tables 3 and 4, reports the mean and standard deviation across 3 random seeds

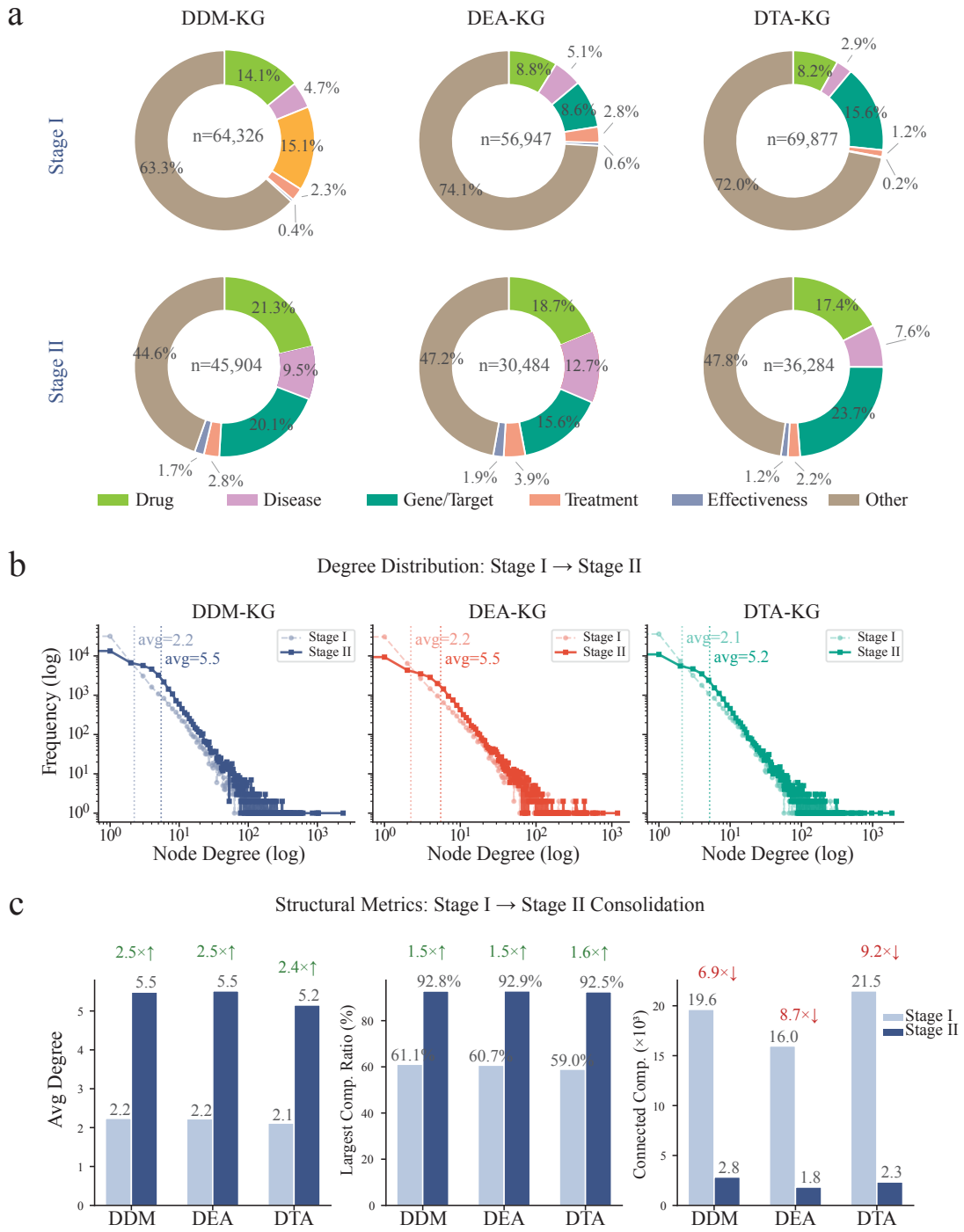


Figure 3: Structural overview of the six UniD³ knowledge graphs across three tasks (DDM, DEA, DTA) and two extraction stages. (a) Entity-type composition. (b) Top relations. (c) Density and degree distribution before and after Stage II.

of Llama-3.3-70B, per-prompt-variant scores under three rephrasings of the instruction $\mathcal{I}_t$ used to

quantify prompt sensitivity, and per-disease-class breakdowns that complement the global metrics

Variant	DDM Cov@3	DEA F1	DTA BERTsc
UniD³ (Mix retriever)	0.8388	0.9883	0.9871
– Stage II re-extraction	0.6968	0.8743	0.9561
– Prompt agreement filter	0.7945	0.9313	0.9742
LightRAG-global only	0.7968	0.8589	0.9677
LightRAG-local only	0.7655	0.8627	0.9418
No retrieval (Llama-3.3 base)	0.4311	0.7762	0.9092

Table 6: Ablation of the four UniD³ components; the released bundle (Appendix F) carries all metrics. Removing Stage II degrades every metric the most, confirming cross-document re-extraction as the largest single source of structural improvement; the prompt agreement filter gives a smaller consistent gain. Both retriever-only ablations sit clearly below Mix, so local and global evidence are complementary. DEA F1 and DTA BERTScore are self-consistency bounded (Table 4); relative gaps across rows remain interpretable.

relation as ineffective (which grows the negative set from 1,152 to 1,446 pairs), moves disease-condition agreement by less than 0.5 pp on both classes. Separately, two licensed clinicians graded a 46-item borderline sample drawn from mid-band patient ratings and UniD³-Data-versus-benchmark disagreements, exactly where the projection is least determinate. The two experts agree at Cohen’s $\kappa = 0.88$; on the 39 items with a decidable consensus, that consensus matches the UniD³-Data label in 87%, and on the mid-band subset in 96% (27 of 28). The residual mismatches are off-label but guideline-supported uses that UniD³-Data conservatively marks ineffective (metformin for PCOS, spironolactone for acne, tadalafil for BPH), a bounded failure mode rather than projection noise.

Ontology grounding of entity identity. Because synonym splits are the single largest residual category, normalised Levenshtein matching (§3.4) is not by itself an adequate identity criterion for biomedical entities. We therefore linked a 500-entity sample per type to canonical public identifiers: 46.4% of drugs anchor to RxNorm, 56.4% of diseases to MeSH and 78.4% of genes to NCBI Gene (no API errors). Grounding repairs real splits, with the 392 anchored gene names collapsing to 96 unique NCBI Gene identifiers. We therefore adopt ontology identifiers as the canonical merge key on top of the Levenshtein step, a post-hoc re-keying through free public lookups that needs no GPU time or re-extraction and improves both graph qual-

ity and the derived splits. The moderate drug and disease rates reflect long-tail, non-canonical surface forms; DrugBank, UMLS and UniProt extend the same key.

H Error Analysis Details

Section 5.7 sketches five recurring failure modes in the Mix-retriever predictions. We expand each category here with one representative example (drug names lightly anonymised where the underlying PubMed source is sparse).

(a) Synonym splits (~ 28% of DDM errors). The pipeline produces *cisplatin-ovarian cancer* and *cisplatinum-ovarian cancer* as separate Stage II edges; the benchmark has only the former. Mitigation: tighter Merge_τ, dictionary-augmented synonym lists.

(b) Indication vs. contraindication (~ 19%). On DrugReviews, a single review may report both efficacy and adverse events. UniD³ prefers the dominant sentence, occasionally producing a *treats* edge for what the benchmark labels a contraindication.

(c) Label granularity mismatch (~ 17%). UniD³ extracts *lung adenocarcinoma*, the benchmark uses the parent MeSH term *lung neoplasm*; the underlying biomedical relation is correct but is scored as a miss under exact-match.

(d) Long-tail entities (~ 22%). For rare diseases and recently approved biologics with ≤ 5 PubMed articles the dual-stage extractor lacks signal. We recommend re-running Stage II once additional literature accumulates.

(e) Plausible repurposing candidates (~ 14%). UniD³ surfaces, for example, *metformin-tuberculosis* as a *treats* edge backed by adjunctive-therapy literature (Singhal et al., 2014). The benchmark does not list this pair, but the supporting PubMed evidence makes it a candidate repurposing hypothesis worth examining rather than a clear-cut error. We argue these should not be counted as errors but as candidate hypotheses.

I Release, Licensing and Reproducibility

Licensing. The construction corpus is the PMC Open Access Subset, whose articles carry Creative Commons or similar licences that permit redistribution and text mining; we redistribute derived triples

and QA items rather than article full text. The released graphs, datasets, benchmark scripts and UniD³-QA code are published under a permissive open licence at the repository above, and each released record carries the licence string of its source article.

Prompt revision identifiers. Every extraction and generation prompt in Appendix A is released under a version tag of the form `task-stage-vN` (for example `ddm-stage2-v3`), and each released record stores the tag of the prompt that produced it, so any record can be traced to the exact prompt revision.

Record-level filtering labels. Records are released with the filter decisions attached rather than silently dropped: each carries its task bucket (DDM/DEA/DTA), the cross-LLM agreement outcome, and language and length filter flags. Consumers can therefore reconstruct any filtering stage, or opt out of it.

One-example rebuild. We ship a script that rebuilds a single worked example end to end, from article retrieval through Stage I and Stage II extraction to the final QA record, using the OMADA-CYCLINE case of Appendix A. It runs in minutes on one GPU and reproduces the released record exactly, letting a reader verify the pipeline without the full 3,530 A100-hour build.

J Compute Resources

All UniD³ extractions and dataset generations were run with Llama-3.3-70B on a single NVIDIA A100-80G GPU and a 4-core Intel Xeon Gold 6142 CPU host. Table 7 reports the per-task token throughput and wall-clock time for each stage; the totals are 127.6M input tokens, 202.7M output tokens, and 3,529.9 A100-hours.

Stage II re-extraction is the dominant cost item ($\approx 59\%$ of total time), as expected from the design of Section 3.4: every entity in $\mathcal{V}_1^{(t)}$ triggers an additional KG-level LLM call, so cost scales with the number of consolidated entities rather than with the chunk count. Translated into commercial pricing at the time of writing, the full pipeline corresponds to a one-off cost of roughly \$8–10K, several orders of magnitude below the cost of comparable manual curation of 157,849 biomedical articles. Inference cost of UniD³-QA under the Mix retriever is dominated by generation on Llama-3.3-70B; we report a wall-clock benchmark for smaller open backbones

(Llama-3.1-8B, Qwen-2.5-32B) alongside the released code.

K KG Quality Discussion

Synonym handling. Drug names exhibit substantial surface variation (brand vs. generic, hyphenation, salt forms, dosage forms). Our fuzzy-match Merge_τ catches the majority of these, but three patterns remain difficult: acronym–expansion pairs such as *5-FU* vs. *fluorouracil*; drug combinations written as “X+Y” vs. “X / Y”; and salt-form variants such as *cisplatin* vs. *cisplatinum*. Manual inspection of 200 randomly sampled drug clusters shows residual synonym splits in $\approx 7\%$ of clusters; these are the dominant source of the synonym-split errors reported in Section 5.7 and Appendix H.

Hallucination spot-check. We sampled 200 Stage II triples and verified each against the sentence-level evidence stored in $\mathcal{E}^{(t)}$. 186 of 200 triples were directly supported by a sentence from at least one cited PubMed article; of the remaining 14, 9 were paraphrases that an expert judged faithful to the source and 5 were genuine hallucinations introduced during Stage II. Hallucination rate at the triple level is therefore $\approx 2.5\%$; this number is the operational ceiling on the trust we ascribe to any single extracted edge, and motivates the human-in-the-loop deployment guideline in Ethical Considerations.

Missing edges. Recall losses come predominantly from articles that mention the drug and the disease only via pronouns or section-cross-references, where the chunk window cannot capture both within a single context, and from emerging drug classes whose PubMed coverage falls below the chunking threshold of Appendix C. Stage II mitigates both effects but does not eliminate them.

L Drug-Centric Exploration of the DEA-KG

To complement the DDM exploration in the main body (Section 5.2, Figure 2), we provide the analogous analysis on the Stage II DEA graph centred on the ANTIBIOTICS node, summarised in Figure 5.

Ego-network. Figure 5(a) shows that the ANTIBIOTICS-centred subgraph spans multiple entity types, including specific antimicrobial agents (*Vancomycin*, *Ciprofloxacin*, *Daptomycin*, *Colistin*, *Gentamicin*), bacterial pathogens (*Staphylococcus*

Task	Stage I (paper-level)			Stage II (KG-level)			Dataset Generation			Total (h)
	In (M)	Out (M)	Hours	In (M)	Out (M)	Hours	In (M)	Out (M)	Hours	
DDM	45.7	29.7	427.2	12.3	43.4	859.2	0.57	4.53	98.3	1,384.7
DEA	35.6	28.1	386.4	8.4	30.1	549.6	0.23	0.97	48.7	984.7
DTA	46.3	31.4	441.6	9.7	33.1	662.4	0.24	1.36	56.5	1,160.5
Total	127.6	89.2	1,255.2	30.4	106.6	2,071.2	1.04	6.86	203.5	3,529.9

Table 7: Token throughput (in millions) and wall-clock time (in hours on a single A100-80G with a 4-core Intel Xeon Gold 6142 CPU) for the UniD³ pipeline. Stage II accounts for $\sim 59\%$ of total compute, consistent with the design that scales the second pass with the number of consolidated entities.

aureus, *MRSA*, *Pseudomonas aeruginosa*, *Klebsiella* / *Acinetobacter baumannii*, *Mycobacterium tuberculosis*), disease conditions (*Bacterial infections*, *Pneumonia*, *Tuberculosis*, *UTI*), and clinically relevant concepts such as *Antibiotic resistance*, *Drug resistance*, *Multidrug resistance* and *Treatment outcomes*. Node size is proportional to degree centrality. The diversity of entity types and edge density show that the DEA task elicits unified relational representations that integrate drug pharmacology, microbial biology, resistance mechanisms and clinical efficacy assessment.

Multi-hop paths. Figure 5(b) traces representative paths originating from ANTIBIOTICS and traversing TREATMENT OUTCOME into the oncology subgraph, where they connect to entities such as *Chemotherapy*, *EGFR*, *Cisplatin*, *Doxorubicin* and *Gemcitabine*. Cross-domain paths of this type are an emergent consequence of the DEA formulation. Efficacy assessment requires reasoning about therapeutic outcomes across disease domains, so the resulting KG naturally captures indirect associations that bridge anti-infective therapy and oncology, which are exactly the kinds of unexpected links that downstream drug repurposing pipelines exploit.

M Drug-Centric Exploration of the DTA-KG

We extend the drug-centric exploration to the Stage II DTA graph centred on *Cisplatin*, summarised in Figure 6.

Ego-network. The Stage II DTA graph centred on *Cisplatin* (Figure 6a) prominently features molecular targets and gene entities (*P53*, *AXL*, *Bcl-2*, *NRF2*, *Abcg2*) alongside cancer types (*Ovarian*, *Lung*, *Glioblastoma*, *Breast*, *Hepatocellular carcinoma*, *Melanoma*, *Medulloblastoma*) and biological processes (*Apoptosis*, *Cell proliferation*,

DNA damage response, *DNA replication*, *Ferroptosis*, *Cell death*). The presence of resistance-related concepts (*Chemoresistance*, *Drug resistance*, *Multidrug resistance*, *Resistance mechanisms*) indicates that the DTA task captures not only direct drug–target interactions but also the broader molecular context of treatment response and failure.

Multi-hop paths. Figure 6(b) traces two primary pathways from *Cisplatin*: one through CANCER TREATMENT and one through *Chemotherapy*, both converging on the CANCER node before diverging into TARGETED THERAPY, COMBINATION THERAPY and the molecular target *EGFR*. Compared with the DEA paths in Figure 5, the DTA paths exhibit greater diversity in intermediate nodes and place stronger emphasis on treatment-modality distinctions (differentiating *Chemotherapy* from *Targeted Therapy* and *Combination Therapy*). This structural difference reflects the design of the DTA task, which requires explicit reasoning about the alignment between a drug’s molecular targets and disease-specific pathway biology.

N Examples of UniD³-Generated Datasets

We illustrate the three sub-datasets with one representative entry per task: Tables 8, 9 and 10 show DDM, DEA and DTA examples respectively. Each example has been verified against the supporting Stage II KG path and the cited PubMed evidence sentences.

O Chatbot Case Studies

Setting. We compared UniD³-QA against the same Llama-3.3-70B backbone without retrieval on 50 PubMedQA-derived questions covering oncology, infectious disease, metabolic disease, cardiovascular disease and CNS disease. Three biomedical reviewers independently chose the better answer between two anonymised, randomly ordered responses, scoring *better*, *tie* or *worse*.

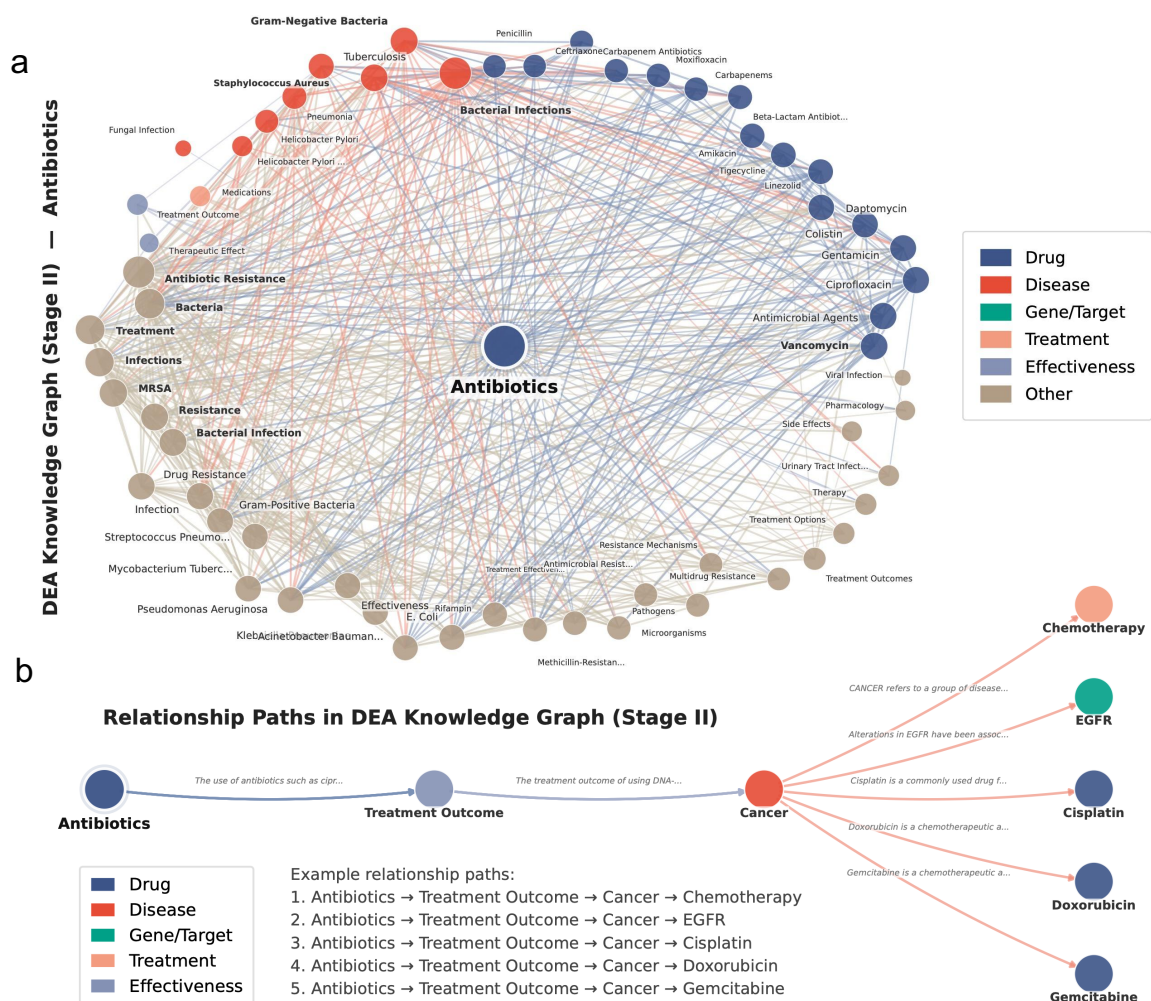


Figure 5: Drug-centric knowledge exploration of the Stage II DEA knowledge graph. (a) Two-hop ego-network around ANTIBIOTICS. (b) Representative multi-hop relation paths from ANTIBIOTICS through TREATMENT OUTCOME into oncology (CANCER, EGFR, CHEMOTHERAPY, CISPLATIN, DOXORUBICIN, GEMCITABINE). Adapted from Supplementary Figure 12 of the original submission.

Aggregate results. UniD³-QA was preferred in 39/50 (78%) cases, tied in 7 and preferred against in 4. Across reviewers, free-text justifications were dominated by “cites the supporting PubMed evidence” and “provides a mechanism chain”, confirming that the practical value of the system is not stronger generation but auditable grounding.

Representative transcript. For the question “Is trastuzumab effective for HER2-positive metastatic breast cancer?”, the base Llama-3.3-70B answers with a generic affirmative paragraph and no source attribution. UniD³-QA returns: a yes/no decision conditioned on the retrieved KG path $\pi = \langle \text{Trastuzumab} \xrightarrow{\text{targets}} \text{HER2} \xrightarrow{\text{overexpressed-in}} \text{HER2+MBC} \xrightarrow{\text{treated-by}} \text{combination chemotherapy} \rangle$; a 3-sentence explanation grounded in the corresponding PubMed sentences (Slamon et al., 2001); and a

confidence label estimated from the number of supporting Stage II edges. Two additional transcripts (cisplatin–ovarian cancer, metformin–tuberculosis) are included in the released artefact.

Failure modes. The four cases in which reviewers preferred the base model involved questions on emerging biologics with thin PubMed coverage; in these cases UniD³-QA correctly abstained but the abstention was judged less informative than the base model’s high-confidence guess. We treat this as a desirable failure mode for a biomedical assistant, but flag it as a UI design point for future work.

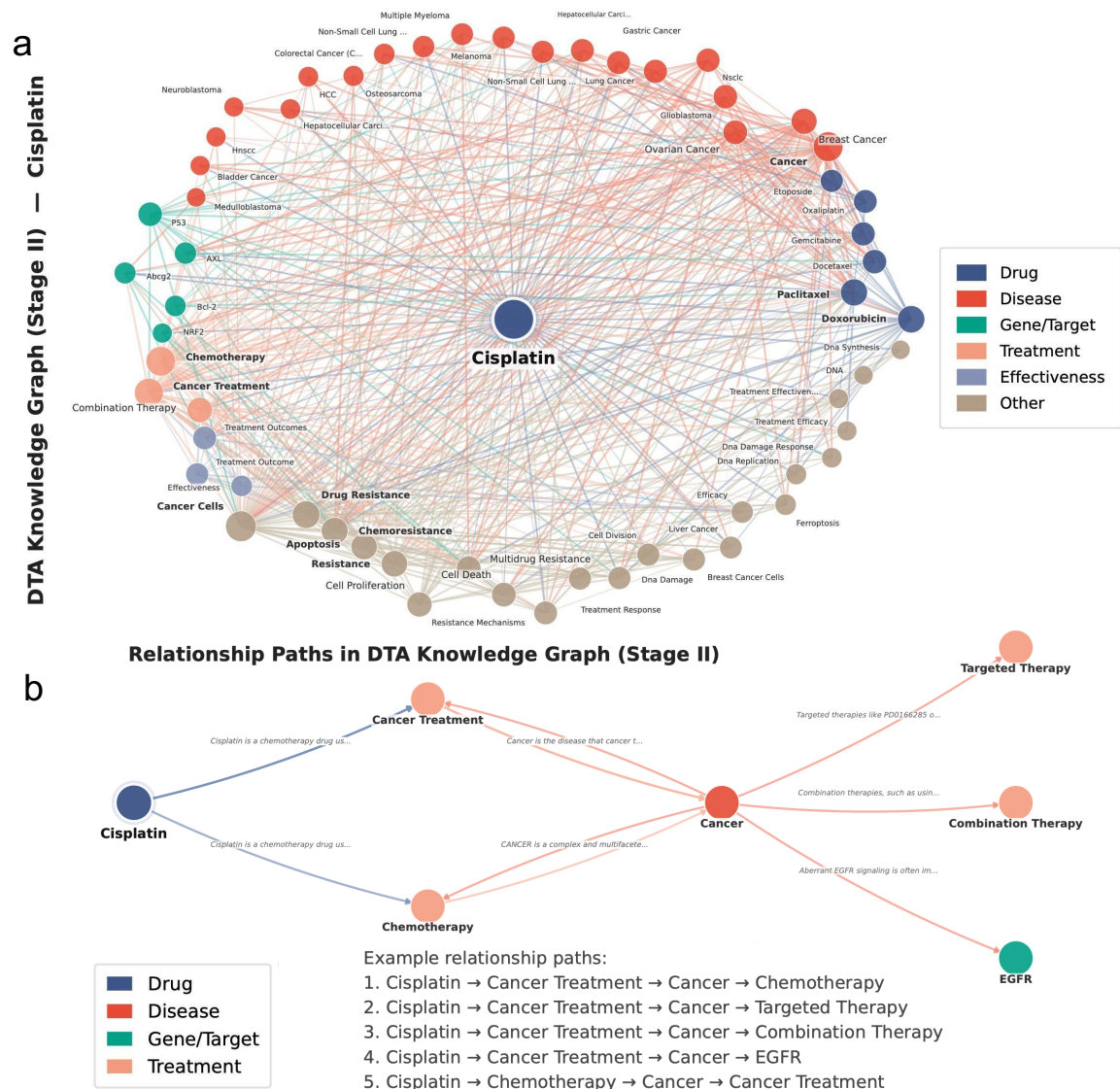


Figure 6: Drug-centric knowledge exploration of the Stage II DTA knowledge graph. (a) Two-hop ego-network around *Cisplatin*, with prominent molecular targets (P53, AXL, Bcl-2, NRF2, Abcg2), cancer types and biological processes (DNA damage response, apoptosis, ferroptosis, drug resistance). (b) Representative multi-hop paths through CANCER TREATMENT and CHEMOTHERAPY into TARGETED THERAPY, COMBINATION THERAPY and *EGFR*. Adapted from Supplementary Figure 13 of the original submission.

Supplementary Table 1: DDM example.

Drug	Lomustine
Disease	Brain tumours (glioblastoma)
Explanation	Lomustine is a nitrosourea alkylating agent that diffuses across the blood–brain barrier and induces DNA cross-links via chloroethyldiazohydroxide intermediates. In glioblastoma, the activity of MGMT-mediated DNA repair determines sensitivity; tumours with low MGMT promoter methylation respond better to lomustine. The induced DNA damage activates the p53 tumour-suppressor pathway, leading to cell-cycle arrest and apoptosis in glioma cells.
Summary	Lomustine belongs to the alkylating agent class. Its primary biochemical activity is DNA cross-linking, and its downstream effects include MGMT-pathway-dependent cytotoxicity and p53-mediated apoptosis.

Table 8: Representative DDM entry. The *Explanation* field provides a disease-contextualised mechanistic rationale; the *Summary* field gives a broader pharmacological characterisation. Adapted from Supplementary Table 1 of the original submission.

Supplementary Table 2: DEA example.

Drug	Abatacept
Disease	Rheumatoid arthritis
Label	Effective
Explanation	Abatacept is a CTLA-4–Ig fusion protein that selectively inhibits T-cell co-stimulation by binding CD80/CD86 on antigen-presenting cells. Multiple randomised controlled trials have demonstrated significant ACR20/50/70 response improvements in RA patients, including those refractory to anti-TNF therapy, and the drug is also approved for juvenile idiopathic arthritis. Evidence for related conditions (psoriatic arthritis, ankylosing spondylitis) is less conclusive, with smaller trials suggesting modest efficacy that does not yet support broad clinical use.

Table 9: Representative DEA entry. The *Label* field stores a categorical efficacy judgment; the *Explanation* synthesises clinical evidence including conflicting findings across disease subtypes. Adapted from Supplementary Table 2 of the original submission.

Supplementary Table 3: DTA example.

Drug	Cobimetinib
Question	How does Cobimetinib’s mechanism of action align with the molecular biology of BRAF V600-mutated melanoma?
Answer	Cobimetinib is a selective allosteric inhibitor of MEK1 and MEK2, the immediate downstream kinases of BRAF in the RAS–RAF–MEK–ERK (MAPK) signalling cascade. In BRAF V600-mutated melanoma, constitutive activation of mutant BRAF drives unregulated MEK–ERK signalling and tumour proliferation. By inhibiting MEK1/2, cobimetinib blocks ERK phosphorylation and re-establishes normal regulation of cell-cycle and apoptotic programs in mutant cells. The mechanism aligns precisely with the dependency of BRAF V600-mutated tumours on sustained MAPK signalling, which is why cobimetinib is clinically combined with the BRAF inhibitor vemurafenib to delay the emergence of MAPK-reactivation–driven resistance.

Table 10: Representative DTA entry. The QA format probes target-level specificity and biological plausibility. Adapted from Supplementary Table 3 of the original submission.